

Cardiorespiratory signature of neonatal sepsis: Development and validation of prediction models
in 3 NICUs

Sherry L. Kausch¹, Jackson G. Brandberg², Jiaying J. Qiu², Aneesha Panda⁴, Alexandra Binai⁴,
Joseph Isler³, Rakesh Sahni³, Zachary A. Vesoulis⁴, J. Randall Moorman², Karen D. Fairchild¹,
Douglas E. Lake², Brynne A. Sullivan^{1*}

1. Department of Pediatrics, University of Virginia School of Medicine, Charlottesville, VA, USA
2. Department of Medicine, University of Virginia School of Medicine, Charlottesville, VA, USA
3. Department of Pediatrics, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA
4. Department of Pediatrics, Washington University in St. Louis School of Medicine, St. Louis, MO, USA

*Corresponding Author: Brynne A. Sullivan
PO Box 800386
Charlottesville, VA 22908
757-291-5985

Abstract

Background: Heart rate characteristics aid early detection of late-onset sepsis (LOS), but respiratory data contain additional signatures of illness due to infection. Predictive models using cardiorespiratory data may improve early sepsis detection. We hypothesized that heart rate (HR) and oxygenation (SpO₂) data contain signatures that improve sepsis risk prediction over HR or demographics alone.

Methods: We analyzed cardiorespiratory data from very low birth weight (VLBW, <1500g) infants admitted to three NICUs. We developed and externally validated four machine learning models to predict LOS using features calculated every 10m: mean, standard deviation, skewness, kurtosis of HR and SpO₂, and cross-correlation. We compared feature importance, discrimination, calibration, and dynamic prediction across models and cohorts. We built models of demographics and HR or SpO₂ features alone for comparison with HR-SpO₂ models.

Results: Performance, feature importance, and calibration were similar among modeling methods. All models had favorable external validation performance. The HR-SpO₂ model performed better than models using either HR or SpO₂ alone. Demographics improved the discrimination of all physiologic data models but dampened dynamic performance.

Conclusions: Cardiorespiratory signatures detect LOS in VLBW infants at 3 NICUs. Demographics risk-stratify, but predictive modeling with both HR and SpO₂ features provides the best dynamic risk prediction.

Introduction

Early detection of late-onset sepsis (LOS, sepsis beyond 3 days of age) reduces mortality and improves outcomes for survivors^{1,2}. Although many infants with LOS exhibit clinical instability, signs and symptoms are nonspecific and occur too late in the course of illness. Signatures of illness are present in physiologic time series data derived from heart rate (HR) and oxygen saturation (SpO₂) monitoring in the early stages of sepsis in premature infants³⁻⁵. We developed and validated algorithms to detect abnormal patterns in continuous HR⁴ and SpO₂⁶ data and in their cross-correlation³.

The heart rate characteristics (HRC) index estimates the risk of imminent sepsis using the standard deviation of RR intervals, sample asymmetry⁷, and sample entropy⁸ to detect decreased HR variability with transient decelerations^{4,9}. A multicenter randomized clinical trial showed HRC display reduced mortality for premature infants^{1,10}.

We have found additional physiological signatures of neonatal illness and explored new data collection, analysis, and modeling methods^{3,5,6}. Respiratory deterioration prompts many sepsis evaluations⁵ in premature infants, and apnea increases during sepsis¹¹ as inflammation affects central control of breathing¹². Thus, we hypothesize that abnormal patterns in pulse oximetry might add to HR characteristics in early detection of sepsis. Here, we used data from three tertiary Neonatal ICUs to develop and validate statistical models combining HR and SpO₂ analytics for sepsis detection in very low birth weight (VLBW, <1500g) infants. We aimed to evaluate multiple cardiorespiratory features, modeling methods, and performance metrics to test the hypothesis that respiratory data contain signatures of illness caused by sepsis and add information to HR characteristics and demographic variables for risk prediction.

When we can determine that signatures of illness are present, we can contemplate using statistical models as bedside predictive tools. Here, we refer to models of this kind as POWS (Pulse Oximetry Warning System).

Methods

Patients

We studied VLBW infants admitted to 3 NICUs: University of Virginia Children's Hospital (NICU 1, 2012-2021), Morgan Stanley Children's Hospital of New York, Columbia University (NICU 2, 2012-2019), and St. Louis Children's Hospital, Washington University School of Medicine (NICU 3, 2016-2021). We report the results of this study in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines^{13,14}.

The Institutional Review Boards of each institution approved the study. We excluded infants with major chromosomal or congenital anomalies and those with no vital sign data collected. We collected demographic and clinical variables from the electronic health record and unit databases.

Sepsis definition

Late-onset septicemia (LOS) is the primary outcome of our modeling. Clinicians at each site reviewed blood cultures obtained after three days of age for infants who met inclusion criteria. We recorded the event as LOS if the blood culture was positive, the infant was treated with at least five days of antibiotics, and the culture was preceded by at least two days with no antibiotics. We excluded negative blood cultures, positive blood cultures obtained within seven days of a prior positive blood culture, and positive blood cultures treated as contaminants (defined as <5 days of antibiotics).

HR and SpO₂ data collection and preprocessing

Continuous HR and SpO₂ data were collected from standard NICU bedside monitors (GE, Philips) using the BedMaster system (Hillrom's Medical Device Integration Solution, Chicago, IL; formerly Excel Medical, Jupiter, FL). Electrocardiogram-derived HR and pulse oximeter-derived SpO₂ and pulse rate were collected at 0.5 Hz at NICU 1 and 2. NICU 3 collected data at 1 Hz and was downsampled to 0.5 Hz to match the other sites. SpO₂ was measured with the default averaging time of 8 seconds, using Masimo technology at NICU 1 and 2 and Nellcor Oximax technology at NICU 3. HR and SpO₂ underwent single-step preprocessing where values representing incontrovertible artifact (zeros) were removed.

Candidate predictors: HR, SpO₂, Demographics

In 10-minute non-overlapping windows, we calculated the mean, standard deviation (SD), skewness, kurtosis of both HR and SpO₂, and their minimum and maximum cross-correlation^{3,5}. We examined the empirical univariate risk of each feature to evaluate individual predictor-risk relationships. We noticed differences in distributions of mean HR between the three sites. Additionally, prior work suggests mean HR does not predict imminent LOS¹⁵. Therefore, we excluded mean HR as a candidate predictor. Furthermore, including mean HR reduced performance across sites.

Model development

Fig 1 provides a schematic overview of the methods, including model development, validation, comparisons, and secondary analyses. Every 10-minute window of raw data was labeled as “control,” “LOS,” or censored. Windows in the 24-hour period preceding the time of positive blood culture were labeled LOS, those falling in the seven days following LOS were censored, and all other windows falling between 72h after birth and NICU discharge were labeled as control. Windows with more than 50% data missing were excluded. All 10-minute windows labeled as LOS and a sub-sample of those labeled as control (one 10-minute window per hour) were used for training. We trained models using data from infants at NICU 1 and tested the models on NICU 2 and 3. Modeling was performed in R (R Foundation for Statistical Computing, Vienna, Austria) and Python (Python Software Foundation, <https://www.python.org/>).

Each model generates the estimated probability of LOS, updated every 10-minutes. To translate model outputs into the fold-increased risk of LOS in the next 24 hours, we divided the estimated probability by the average probability of LOS. We evaluated different smoothing windows that averaged the 10-minute model outputs over different time periods, ranging from the preceding 1 to 12 hours. We determined an optimum smoothing window by balancing the rise in area under the receiver operating characteristic (AUC) against the decrease in the average risk score for LOS. After smoothing, model outputs were sampled hourly. Finally, hourly, smoothed model predictions were recalibrated using the sigmoid function.

To determine importance, features were permuted to calculate the loss in AUC and ranked from highest to lowest. We added ranks across models to estimate the overall feature importance.

Modeling methods

Four modeling strategies were evaluated, including logistic regression, a neural network, an ensemble method (extreme gradient boosting classifier or XGBoost), and random forest (Fig 1).

For model #1, we built an 18-dimensional logistic regression model by including one additional non-linear term^{16,17} for each of the nine features. We adjusted for repeated measures using the Huber-White method¹⁸. We removed minimum cross-correlation, SD SpO₂, and non-linear effects with p-values >0.05. The result was an 11-dimensional model.

For model #2, we developed a neural network model with four hidden layers. These layers had 512, 256, 128, and 64 hidden neurons, respectively. Dropout layers at a rate of 0.3 were added as a regularization measure between each hidden layer¹⁹. Leaky ReLU was used as the activation function in the hidden layers and a sigmoid activation function was used in the last layer²⁰. The model was trained using binary cross-entropy as the loss function and Adam as the optimizer with a learning rate of 0.001. An early stopping callback was used to avoid overfitting which monitored the loss with a patience of 10.

Model #3 was an Extreme Gradient Boosting (XGBoost) classifier²¹. We trained multiple models on a parameter grid to identify optimal hyperparameters for learning rate, alpha (l1 regularization), lambda (l2 regularization), and max depth based on the highest validation AUC on NICU 2 after training on NICU 1. The model used binary logistic as its objective function.

Model #4 was a random forest model with 800 classification trees and the square root of the number of features were sampled as candidates at each split²². The output of the model was the fraction of trees that classified an outcome as an event.

Model validation

We validated models on unseen data from two independent sites. We trained models on NICU 1, the largest and most complete data set, and externally validated on data from the other sites. All subsequent

analyses of model validation and performance used the models trained on data from NICU 1. We assessed model discrimination by AUC with confidence intervals based on 200 bootstrap runs.

We evaluated model calibration by plotting the predicted vs. observed relative risk of LOS. We assessed model performance in the critical period near the time of LOS diagnosis by examining the time course of the average relative risk. We identified a rise in risk by performing a sign-rank test with the null hypothesis that hourly risk estimates are equal to risk estimates from the same patient 24 hours prior with a statistical significance threshold of $p < 0.05$.

Sensitivity analyses

We assessed the dynamic change in the model output near the time of other “sepsis-like” events, including negative blood cultures diagnosed as clinical sepsis or necrotizing enterocolitis (NEC) without bacteremia. Because mechanical ventilation can alter HR and SpO₂ patterns^{3,23,24}, we also examined differences in models’ predicted probability of sepsis based on ventilator status. Given the potential confounding of coagulase-negative staphylococcus (CONS),^{25,26} we conducted a subgroup analysis to assess model performance in CONS versus non-CONS positive blood cultures.

We tested whether models performed equivalently using features derived only from pulse-oximetry data. While ECG and pulse oximetry monitoring are standard, resource-poor settings might benefit from an algorithm that operates from an oximeter alone. Therefore, we tested model performance using features derived from pulse rate rather than the features derived from ECG HR.

Model comparisons

We hypothesized that HR, SpO₂, and static demographic variables are independent predictors of LOS. To quantify the added value, we built three additional models. First, we trained a demographics model that included birth weight, sex, and chronological age as baseline risk factors^{4,15,27,28}. Second, we created a HR model that contained SD, skewness, and kurtosis. Third, we developed an SpO₂ model that contained mean, SD, skewness, and kurtosis. For comparison to a validated measure, we also calculated the HRC index^{1,4}.

At a range of thresholds, we calculated the number of alerts per patient day, where an alert had to start within the day preceding the clinical diagnosis of sepsis and could only fire once per day. We excluded alerts in the seven days following a sepsis event.

Results

Participants

We studied 3,151 VLBW infants at the three sites. Vital sign data were available for 2494/3151 (79%) infants and 302 of 390 episodes of LOS (77%). Table 1 shows the characteristics of the study population. They are typical of a VLBW cohort, with minor inter-center differences. The median age of LOS was 16 days (IQR: 8 - 32). As expected, infants with LOS had lower gestational age and birth weight.

Model development

We trained models on 923 infants from NICU 1. The nine candidate predictors and their empirical relationship with the outcome of sepsis are shown in S1 Fig. The retained features used in the logistic regression model were mean SpO₂, SD HR, skewness of HR, skewness of SpO₂, kurtosis of HR, kurtosis of SpO₂, and maximum positive value of the HR-SpO₂ cross-correlation. Table S2 presents model coefficients.

After models were trained, we engaged in three post-processing steps. First, we evaluated different smoothing windows that averaged the 10-minute model outputs over different time periods. The optimum smoothing window balanced the rise in AUC against the decrease in the average risk score for LOS. The optimal smoothing window was four hours (S2 Fig). After smoothing all model outputs over 4 hours, we sampled the model outputs hourly (rather than every 10-minutes). Finally, we recalibrated prior outputs prior to performance evaluation and validation.

Model performance

All modeling methods discriminated sepsis windows from control windows with good performance (AUCs > 0.8, Table 2), and external validation showed only small performance diminishment. Additionally, model performance was similar whether using ECG-derived HR or pulse rates from

oximetry (AUC loss -0.001, -0.006, -0.013 at NICU 1, 2, and 3, respectively). There was high and consistent calibration (Fig 2). When the observed risk of sepsis is less than 1, all models slightly underpredicted risk at NICU 2 and overpredicted at NICU 3.

Fig 3 plots the average fold-increase in each risk model as a function of time to LOS at each NICU. A fold-increase of 1 indicates no greater risk for sepsis than the baseline (0.26%); a fold-increase of 2 would indicate twice the average daily risk (0.52%). The relative risk predicted by the logistic regression model increased by 92% in the 24 hours before sepsis (2.5-fold to 4.8-fold), and the predicted risk calculated in the XGBoost model increased by 150%, from 3.2-fold to 8.0-fold. Predicted risk values deviated significantly from baseline from 23 to 24 hours before the blood culture (Fig 3).

Feature importance

Fig 4 shows feature importance in the logistic regression (A), neural net (B), XGBoost (C), and random forest (D) sepsis prediction models. Features are ordered by decrease in AUC introduced by permuting each feature. Of note, seven features were included in the logistic regression model, while nine features were used in the remaining models. The features of greatest importance, identified by adding the ranks of the features across models, were skewness of HR, SD of HR, kurtosis of SpO₂, and maximum cross-correlation. The random forest model identified kurtosis of SpO₂ as the feature of greatest importance, while the remaining models ranked it as less important than skewness of HR and SD of HR.

Do HR-SpO₂ models add information to HR and demographics?

Having established that there is a cardiorespiratory signature of neonatal sepsis, we evaluated the potential clinical utility of the statistical models. We chose the logistic regression model, and we will call it POWS, for Pulse Oximetry Warning System.

We compared POWS to models with features derived from only HR data, only SpO₂ data, or from only demographics of chronological age, sex, and birth weight. The parameters for each of these models are shown in supplementary tables.

Table 3 shows the results of model discrimination for LOS in training and testing data. The POWS model performed better than models using only HR-derived (AUC, 0.82 vs 0.80 NICU 1) or only SpO₂-

derived features (AUC, 0.82 vs 0.71, NICU 1) or only demographics (AUC 0.82 vs 0.78 NICU 1). As expected, the demographics model provides only near-static information (Figure 3).

To further understand how combining HR and SpO₂ data features in prediction models adds to HR characteristics alone, we compared POWS to the HRC index, a logistic regression model that uses only features derived from HR data⁴. Although POWS and HRC both utilize HR-derived features, POWS uses every two-second HR data to calculate HR SD, skewness, and kurtosis, while the HRC index uses inter-beat-intervals to calculate HR SD, sample asymmetry, and sample entropy. Despite these differences, POWS performed similarly to the HRC index in the NICU 1 cohort (HRC index AUC 0.795, CI 0.798 - 0.803). Fig 5 displays predictiveness curves of the HRC index and POWS model. POWS fits the data better for prediction of LOS in the NICU 1 cohort.

Sensitivity and alarm rates

Examining the sensitivity of the models across a range of thresholds, based on the number of daily alarms in a 50-bed unit, demonstrates that POWS has a perceptibly higher sensitivity than the models using HR or SpO₂ features alone (Fig 6, Fig S5). The low sensitivity of the demographics model again speaks to the limitations of static variables for continuous risk prediction. Adding demographic features to POWS decreases the sensitivity across the range of alert rates and therefore limits its utility as an early warning score.

We examined lift charts as a way to visually compare the ability of each model to detect LOS events (Fig S6). Lift is the ratio of the percent of events captured at a given percentile of data to the random expectation of the percent of events captured. POWS had a higher lift than HR-only models, and both had a greater lift than the SpO₂-only or demographics-only models. We performed this analysis to demonstrate the added value of combined cardiorespiratory modeling to detect sepsis within the highest risk segments of data.

Model performance in other clinical contexts

We trained our models on blood-culture positive sepsis, but assessed model performance around the time of clinical deterioration with other diagnoses and stratified by factors that denote illness severity.

First, we examined the dynamic risk estimates of POWS near the time of blood cultures for clinical sepsis (negative blood culture treated with antibiotics for at least five days due to clinical illness) and the time of blood cultures for confirmed NEC without bacteremia. Fig S3 shows a steep rise in risk in the hours preceding cultures for clinical sepsis and NEC, with a slower decrease in risk in the 48 hours following cultures.

Fig S4 shows model performance near LOS events, stratified by ventilator status at the time of sepsis and by causative organism, grouped as CONS vs. non-CONS. On average, LOS episodes on a ventilator and those caused by non-CONS bacteremia had higher risk scores before, during, and after diagnosis of LOS.

Discussion

We show that cardiorespiratory signatures provide important information for early warning of LOS in VLBW infants. Prior work identified abnormal heart rate characteristics and increased HR-SpO₂ cross-correlation as physiomarkers of illness due to sepsis^{3,5,29}. Here, we report that more information exists in the data at hand. Using machine learning modeling on multicenter data, we tested the hypothesis that HR and SpO₂ data contain patterns that contribute independent, additive information for LOS detection. In doing so, we developed and validated cardiorespiratory models that predict an increased risk of sepsis before clinical diagnosis, up to 24 hours before the time of blood culture. External validation confirmed that these signatures may be generalized for sepsis detection at centers with variable patient populations, monitoring equipment, and practice patterns. Additionally, we found that static demographic variables help to risk-stratify infants at baseline but dampen the dynamic risk prediction.

A cardiorespiratory signature of neonatal sepsis

Previously, we showed that a signature of sepsis exists in the HR signal^{9,30}. Decreased HR variability with transient HR decelerations was recognized as similar to patterns of fetal distress that also served as a physiologic biomarker of sepsis in premature infants³¹. A decade of prior work translated abnormal HR patterns into mathematical algorithms^{7,8,32}, produced a validated predictive model for LOS in VLBW infants³², known as the HRC index or HeRO score, and demonstrated reduced mortality in a multicenter

randomized trial¹. The past work exemplified a pathway for the translation of ideas about signatures of illness in continuous cardiorespiratory monitoring data into bedside tools for clinicians and patients.³³

More recently, our group studied control of ventilation in premature infants and recognized that the HR and SpO₂ signals correlate when episodes of apnea or periodic breathing lead to decreases in both HR (bradycardia) and SpO₂ (desaturation)^{3,34}. Addition of the HR-SpO₂ cross-correlation coefficient improved the performance of statistical models to detect sepsis over HRC monitoring alone in a two-NICU cohort. We found HR-SpO₂ cross-correlation to be the best individual feature to discriminate LOS vs. sepsis-ruled out events.⁵ We also compared vital signs across our three collaborating sites and found clinically trivial though statistically significant differences in HR and SpO₂³⁵.

Respiratory deterioration prompts a majority of the sepsis evaluations in VLBW infants. Features in the POWS model quantify HR and SpO₂ signals and their interaction, thereby capturing reduced variability, decelerations, desaturations, and other more subtle components of a signature of cardiorespiratory deterioration due to sepsis.

These findings support the idea that analysis of respiratory data, in addition to heart rate data, can enhance early detection of subacute but potentially catastrophic illnesses such as sepsis.

Machine learning methods yield similar results

Interestingly, the choice of machine learning methods did not significantly impact performance. Logistic regression performed comparably to more complex methods. Unlike our prior work,^{3,5} here we used cubic splines to account for non-linear relationships of predictors. We hypothesize that this resulted in similar performance of logistic regression compared with other machine learning methods that allow for non-linear relationships. Given equal performance but better explainability, we used the logistic regression approach for the remainder of the analysis. Our finding is matched by others in the literature; previous analyses of multiple machine learning methods for detecting LOS have also found similar performance across modeling techniques^{36,37}.

Comparison with other work

Other groups have used machine learning methods and high-resolution cardiorespiratory data to develop models predicting LOS. Researchers in the Netherlands developed novel algorithms to detect reduced infant motion^{38,39} and demonstrated improved LOS prediction when combined with features measuring heart rate, SpO₂, and respiratory rate⁴⁰. We also find that combining multiple physiological features improves LOS prediction and extended the analysis to larger and external cohort. Other studies have included respiratory rate (RR)^{39,41}, but our prior work has shown RR-derived features to be less predictive in sepsis models³.

Studies have repeatedly shown clinical variables and laboratory values to add to vital sign data for sepsis prediction^{5,36,41-43}. We found that demographic variables improve model AUCs, but dampen the dynamic performance of a continuous sepsis risk model. Physiologic data is continuous, objective, and contains the earliest signs of the inflammatory response to infection via the autonomic nervous system⁴⁴. While including demographic variables or clinician-initiated laboratory data may result in higher AUC, it may limit the lead time for early detection of sub-acute deterioration due to sepsis.

Limitations

We captured a large number of infants and events, but some infants and events were excluded due to missing data. At NICU 2, the data were missing at random, while at NICU 3 the data retrieval process favored the smaller, sicker infants (less likely to change rooms, interrupting the flow of data). Prospective analysis with comprehensive data capture is needed to test models and risk trajectories over time and episodes in a multicenter cohort of premature infants.

Definitions of neonatal sepsis in premature infants vary widely among published studies^{45,46} due to challenges in developing a consensus definition for sepsis in this unique population⁴⁷. We chose to model events of culture-proven sepsis, knowing that this definition, while unequivocal, also excludes many clinically important events associated with a negative blood culture. The heterogeneity of non-culture proven events likely impacts model development⁴¹ and, therefore, performance on new data. Nevertheless, our models trained exclusively on culture-positive events displayed a rise in predicted risk

near the time of negative blood cultures diagnosed as clinical sepsis or NEC, indicating utility for detecting deterioration associated with sepsis-like events.

Clinical implications and future directions

The goal of this work is to optimize a physiology-based sepsis early warning system to improve outcomes by bringing the clinician to the bedside of the right patient at the right time, even when resources are limited to pulse oximetry. All too often, LOS diagnosis occurs after the infection progresses to an advanced phase of systemic inflammation, organ dysfunction, and shock. Our results show that cardiorespiratory predictive monitoring can detect a sub-clinical prodrome in HR and SpO₂ data with superior discrimination and sensitivity compared to analytics from either signal alone or demographic risk factors. External validation of POWS in two geographically distinct cohorts indicates that the signatures of sepsis may be general and are not greatly impacted by center-specific practice patterns or equipment. The ultimate test for validating our models and findings will be a prospective, multicenter clinical trial to measure the impact on clinical care and outcomes.

Conclusion

A cardiorespiratory early warning score, analyzing heart rate from electrocardiogram or pulse oximetry together with SpO₂, predicts late-onset sepsis diagnosis within 24h across multiple NICUs and detects sepsis better than heart rate characteristics or demographics alone.

Data Availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Moorman JR, Carlo WA, Kattwinkel J, Schelonka RL, Porcelli PJ, Navarrete CT, et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *J Pediatr*. 2011 Dec;159(6):900–6.e1.
2. King WE, Carlo WA, O’Shea TM, Schelonka RL, HRC neurodevelopmental follow-up investigators. Heart rate characteristics monitoring and reduction in mortality or neurodevelopmental impairment in extremely low birthweight infants with sepsis. *Early Hum Dev*. 2021 Aug;159:105419.
3. Fairchild KD, Lake DE, Kattwinkel J, Moorman JR, Bateman DA, Grieve PG, et al. Vital signs and their cross-correlation in sepsis and NEC: a study of 1,065 very-low-birth-weight infants in two NICUs. *Pediatr Res*. 2017 Feb;81(2):315–321.
4. Griffin MP, O’Shea TM, Bissonette EA, Harrell FE, Lake DE, Moorman JR. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. *Pediatr Res*. 2003 Jun;53(6):920–926.
5. Sullivan BA, Nagraj VP, Berry KL, Fleiss N, Rambhia A, Kumar R, et al. Clinical and vital sign changes associated with late-onset sepsis in very low birth weight infants at 3 NICUs. *J Neonatal Perinatal Med*. 2021;14(4):553–561.
6. Sullivan BA, Wallman-Stokes A, Isler J, Sahni R, Moorman JR, Fairchild KD, et al. Early pulse oximetry data improves prediction of death and adverse outcomes in a two-center cohort of very low birth weight infants. *Am J Perinatol*. 2018 May 28;35(13):1331–1338.
7. Kovatchev BP, Farhy LS, Cao H, Griffin MP, Lake DE, Moorman JR. Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome. *Pediatr Res*. 2003 Dec;54(6):892–898.
8. Lake DE, Richman JS, Griffin MP, Moorman JR. Sample entropy analysis of neonatal heart rate variability. *Am J Physiol Regul Integr Comp Physiol*. 2002 Sep;283(3):R789–97.
9. Griffin MP, Lake DE, Bissonette EA, Harrell FE, O’Shea TM, Moorman JR. Heart rate characteristics: novel physiometers to predict neonatal infection and death. *Pediatrics*. 2005 Nov;116(5):1070–1074.
10. Fairchild KD, Schelonka RL, Kaufman DA, Carlo WA, Kattwinkel J, Porcelli PJ, et al. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatr Res*. 2013 Nov;74(5):570–575.
11. Fairchild K, Mohr M, Paget-Brown A, Tabacaru C, Lake D, Delos J, et al. Clinical associations of immature breathing in preterm infants: part 1-central apnea. *Pediatr Res*. 2016 Jul;80(1):21–27.
12. Herlenius E. An inflammatory pathway to apnea and autonomic dysregulation. *Respir Physiol Neurobiol*. 2011 Sep 30;178(3):449–457.
13. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015 Jan 6;162(1):55–63.
14. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015 Jan 6;162(1):W1–73.

15. Lake DE, Fairchild KD, Kattwinkel J, Moorman JR. Reply to: Heart rate predicts sepsis. *J Pediatr*. 2012 Oct;161(4):770–771.
16. CRAN - Package rms [Internet]. [cited 2022 Mar 15]. Available from: <https://cran.r-project.org/web/packages/rms/index.html>
17. Harrell FE. rms: Regression Modeling Strategies. 2015;4.3-0.
18. Harrell , FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Cham: Springer International Publishing; 2015.
19. Srivastava N, Hinton G, Krizhevsky A. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine* 2014;
20. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015 May 28;521(7553):436–444.
21. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD' '16*. New York, New York, USA: ACM Press; 2016. p. 785–794.
22. Breiman L. *Random Forests*. Springer Science and Business Media LLC. 2001;
23. van Ravenswaaij-Arts CM, Hopman JC, Kollée LA, Stoelinga GB, van Geijn HP. The influence of artificial ventilation on heart rate variability in very preterm infants. *Pediatr Res*. 1995 Jan;37(1):124–130.
24. Di Fiore JM, Poets CF, Gauda E, Martin RJ, MacFarlane P. Cardiorespiratory events in preterm infants: etiology and monitoring technologies. *J Perinatol*. 2016 Mar;36(3):165–171.
25. Jean-Baptiste N, Benjamin DK, Cohen-Wolkowicz M, Fowler VG, Laughon M, Clark RH, et al. Coagulase-negative staphylococcal infections in the neonatal intensive care unit. *Infect Control Hosp Epidemiol*. 2011 Jul;32(7):679–686.
26. Cantey JB, Anderson KR, Kalagiri RR, Mallett LH. Morbidity and mortality of coagulase-negative staphylococcal sepsis in very-low-birth-weight infants. *World J Pediatr*. 2018 Mar 13;14(3):269–273.
27. Downey LC, Smith PB, Benjamin DK. Risk factors and prevention of late-onset sepsis in premature infants. *Early Hum Dev*. 2010 Jul;86 Suppl 1:7–12.
28. Dong Y, Speer CP. Late-onset neonatal sepsis: recent developments. *Arch Dis Child Fetal Neonatal Ed*. 2015 May;100(3):F257–63.
29. Sullivan BA, Fairchild KD. Predictive monitoring for sepsis and necrotizing enterocolitis to prevent shock. *Semin Fetal Neonatal Med*. 2015 Aug;20(4):255–261.
30. Lake DE, Fairchild KD, Moorman JR. Complex signals bioinformatics: evaluation of heart rate characteristics monitoring as a novel risk marker for neonatal sepsis. *J Clin Monit Comput*. 2014 Aug;28(4):329–339.
31. Griffin MP, Moorman JR. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *Pediatrics*. 2001 Jan;107(1):97–104.
32. Lake DE, Griffin MP, Moorman JR. New mathematical thinking about fetal heart rate characteristics. *Pediatr Res*. 2003 Jun;53(6):889–890.
33. Moorman JR. The principles of whole-hospital predictive analytics monitoring for clinical medicine originated in the neonatal ICU. *npj Digital Med*. 2022 Mar 31;5(1):41.

34. Fairchild KD, Lake DE. Cross-Correlation of Heart Rate and Oxygen Saturation in Very Low Birthweight Infants: Association with Apnea and Adverse Events. *Am J Perinatol*. 2018 Apr;35(5):463–469.
35. Zimmet AM, Sullivan BA, Fairchild KD, Moorman JR, Isler JR, Wallman-Stokes AW, et al. Vital sign metrics of VLBW infants in three NICUs: implications for predictive algorithms. *Pediatr Res*. 2021 Mar 25;
36. Song W, Jung SY, Baek H, Choi CW, Jung YH, Yoo S. A Predictive Model Based on Machine Learning for the Early Detection of Late-Onset Neonatal Sepsis: Development and Observational Study. *JMIR Med Inform*. 2020 Jul 31;8(7):e15965.
37. Masino AJ, Harris MC, Forsyth D, Ostapenko S, Srinivasan L, Bonafide CP, et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PLoS One*. 2019 Feb 22;14(2):e0212665.
38. Joshi R, Bierling BL, Long X, Weijers J, Feijs L, Van Pul C, et al. A Ballistographic Approach for Continuous and Non-Obtrusive Monitoring of Movement in Neonates. *IEEE J Transl Eng Health Med*. 2018 Oct 12;6:2700809.
39. Joshi R, Kommers D, Oosterwijk L, Feijs L, van Pul C, Andriessen P. Predicting Neonatal Sepsis Using Features of Heart Rate Variability, Respiratory Characteristics, and ECG-Derived Estimates of Infant Motion. *IEEE J Biomed Health Inform*. 2020;24(3):681–692.
40. Cabrera-Quiros L, Kommers D, Wolvers MK, Oosterwijk L, Arents N, van der Sluijs-Bens J, et al. Prediction of Late-Onset Sepsis in Preterm Infants Using Monitoring Signals and Machine Learning. *Crit Care Explor*. 2021 Jan 27;3(1):e0302.
41. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *J Am Med Inform Assoc*. 2014 Apr;21(2):326–336.
42. Griffin MP, Lake DE, Moorman JR. Heart rate characteristics and laboratory tests in neonatal sepsis. *Pediatrics*. 2005 Apr;115(4):937–941.
43. Griffin MP, Lake DE, O’Shea TM, Moorman JR. Heart rate characteristics and clinical signs in neonatal sepsis. *Pediatr Res*. 2007 Feb;61(2):222–227.
44. Sullivan BA, Fairchild KD. Vital signs as physiomarkers of neonatal sepsis. *Pediatr Res*. 2022 Jan;91(2):273–282.
45. Henry CJ, Semova G, Barnes E, Cotter I, Devers T, Rafee A, et al. Neonatal sepsis: a systematic review of core outcomes from randomised clinical trials. *Pediatr Res*. 2022 Mar;91(4):735–742.
46. Hayes R, Hartnett J, Semova G, Murray C, Murphy K, Carroll L, et al. Neonatal sepsis definitions from randomised clinical trials. *Pediatr Res*. 2021 Nov 6;
47. McGovern M, Giannoni E, Kuester H, Turner MA, van den Hoogen A, Bliss JM, et al. Challenges in developing a consensus definition of neonatal sepsis. *Pediatr Res*. 2020 Jul;88(1):14–26.

Funding Support: We acknowledge the following grants for funding the work presented in this manuscript:

K23 HD097254 [PI: B Sullivan]; R01 HD092071 [Co-PIs KD Fairchild & JR Moorman, Co-I DE Lake]

K23NS111086 [PI: Z Vesoulis]

Author Contributions: SK, BS, KF, DL, RS, ZV, and JRM have made substantial contributions to the conception or design of the work; SK, JQ, JB, AP, AB, and JI made substantial contributions to the acquisition, analysis, or interpretation of data; SK and BS drafted the work and all other authors have substantively revised it. All authors have approved the submitted version. All authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Competing Interests statement: Some authors have financial conflicts of interest. JRM and DEL own stock in Medical Prediction Sciences Corporation. JRM is a consultant for Nihon Kohden Digital Health Solutions. ZAV is a consultant for Medtronic. All other authors have no financial conflicts to disclose. No authors have any non-financial conflicts of interest to disclose.

Consent Statement: This study was approved by the IRB at each site with waiver of consent.

Figure Legends:

Fig 1. A Schematic Overview of Methods. From top to bottom, we processed the raw signals, sampled at 0.5 Hz, by calculating HR and SpO₂ features every ten minutes. Each 10-minute window, from 72 hours after birth until NICU discharge or death, was labeled as late-onset sepsis (LOS), control, or removed as a blackout period window. Data from NICU 1 were used to train four machine learning models. Before external validation on data at NICU 2 & 3, post-processing steps included smoothing the 10-minute model outputs over 4 hours and recalibrating. Metrics used for external validation included discrimination by AUC, calibration, and plotting the average relative risk over the 48 hours preceding sepsis to look for a dynamic rise from baseline near the time of diagnosis by blood culture.

Fig 2. Calibration plots. Calibration of each POWS model for (A) NICU 1, (B) NICU 2, and (C) NICU 3. Model outputs are smoothed over 4 hours, sampled hourly, and recalibrated after smoothing. Predicted risk relative to average is on the abscissa and observed risk relative to average is on the ordinate. Each point represents one decile of predicted risk. The line of identity is shown as a dashed line. LR = logistic regression, NN = neural network, XG = XGBoost, RF = random forest

Fig 3. The Average Relative Risk of Sepsis The average relative risk of sepsis as predicted by each model as a function of the time to event in hours. Panels show the results of each model at (A) NICU 1, (B) NICU 2, and (C) NICU 3. Results are shown for the four POWS models and the demographic-only model (in gray). Black crosses indicate times where the model outputs are significantly higher ($p < 0.05$) than outputs from the same patient 24 h prior. LR = logistic regression, NN = neural network, XG = XGBoost, RF = random forest

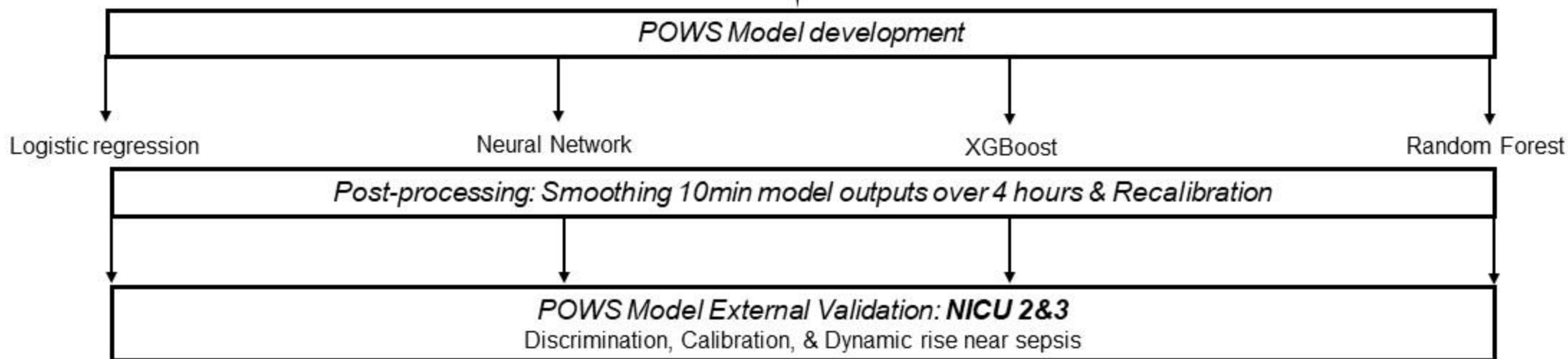
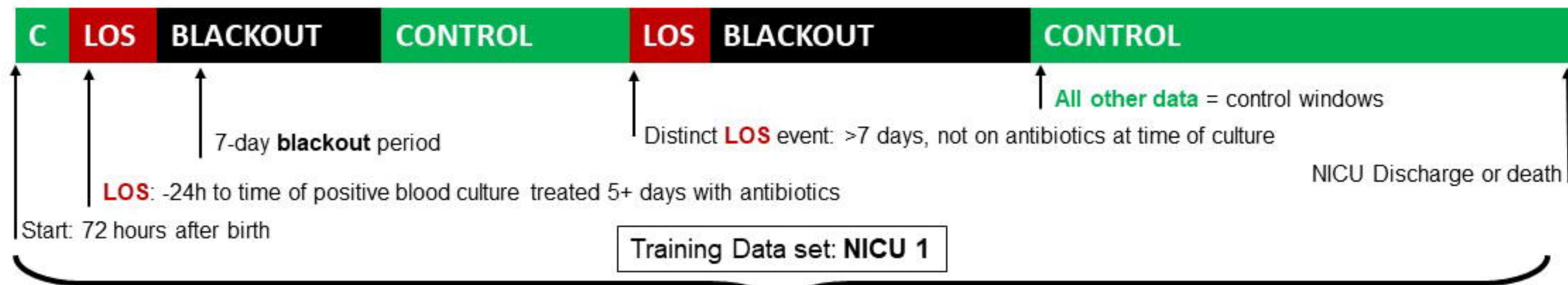
Fig 4. Variable importance plots. Variable importance plots for components of the logistic regression (LR), neural net (NN), XG Boost (XG), and random forest (RF) sepsis prediction

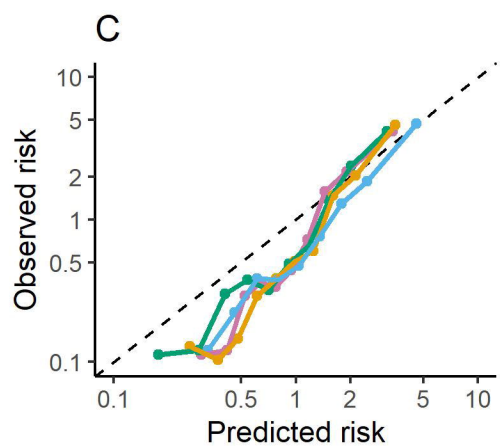
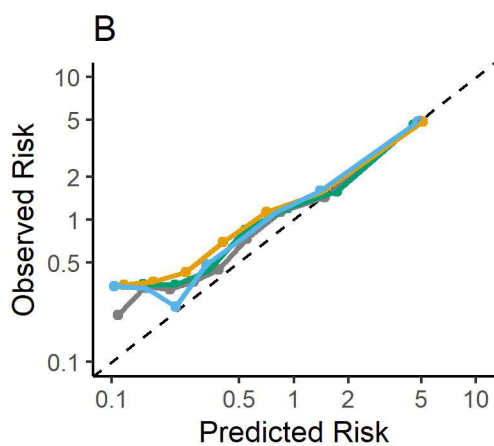
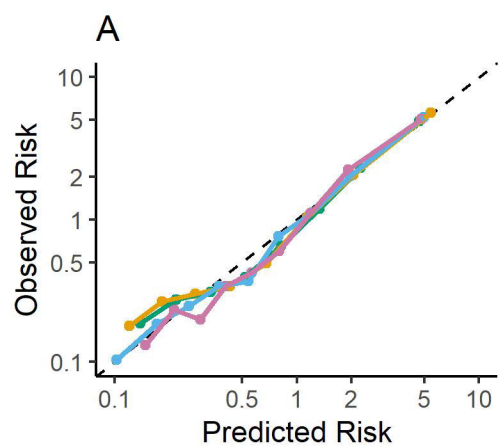
models. Features are ordered by decreasing AUC loss introduced by permuting the values of each feature. Summing the rank of each feature across all four models gives the following overall rank to each feature: 1) Skewness HR, 2) SD HR, 3) Kurtosis SpO₂, 4) Max XC, 5) Mean SpO₂, 6) Kurtosis HR, 7) Skewness SpO₂, 8) Min XC, 9) SD SpO₂.

Fig 5. Predictiveness Curves. Predictiveness curves for HRC index score and POWS LR score in estimating sepsis risk. The solid lines show risk scores arrayed from smallest to largest. The circles are the observed fold in-increase in risk of sepsis. Blue represents the HRC index, and red represents the POWS LR scores.

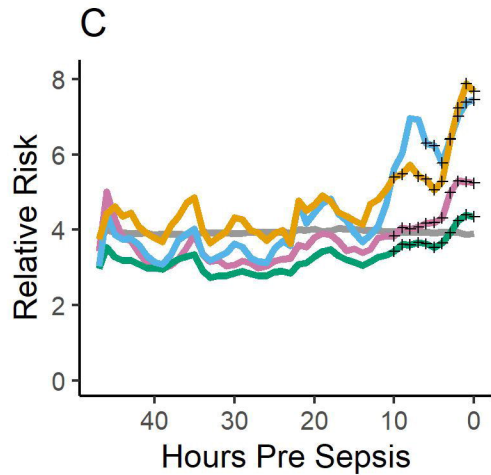
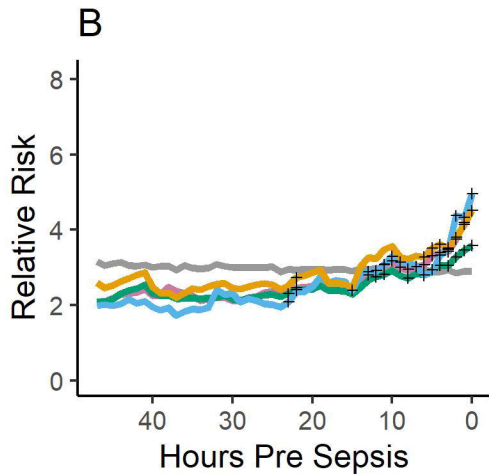
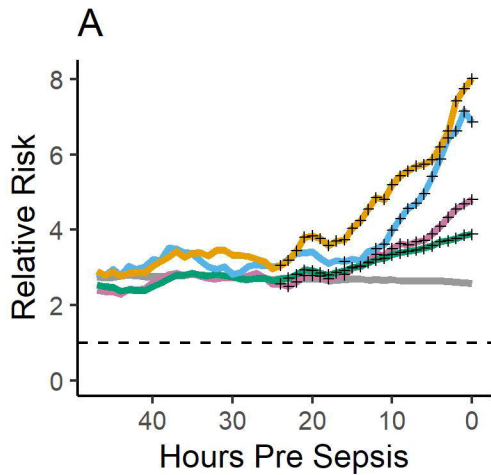
Fig 6. Evaluating the Sensitivity of Models across a Range of Thresholds. We selected a range of thresholds and then calculated the number of alerts per day and required an alert to start within the day preceding the clinical diagnosis of sepsis. We defined alerts as daily threshold crossings. We excluded alerts in the seven days following a sepsis event. The y-axis displays the percent of sepsis events detected when allowing for different numbers of alerts per day. The HRC index (not displayed) performs similarly to the HR-only model.

Signal processing: 0.5 Hz	ECG-derived HR	Pulse oximetry-derived SpO2
Feature extraction: Calculated every 10 minutes	Mean	Mean
	SD	SD
	Skewness	Skewness
	Kurtosis	Kurtosis
	Max & Min HR-SpO2 Cross-correlation	



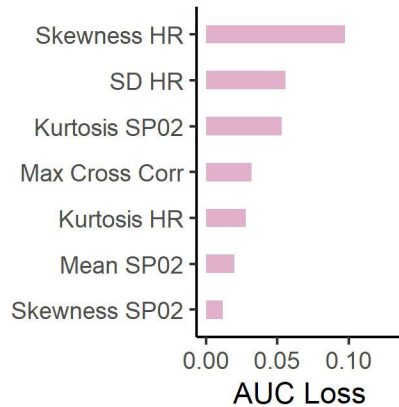


—●— LR —●— NN —●— XG —●— RF

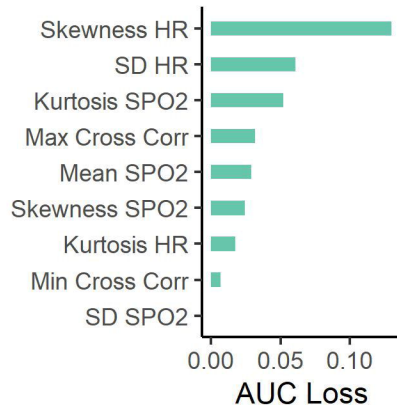


model LR NN XG RF Demographic

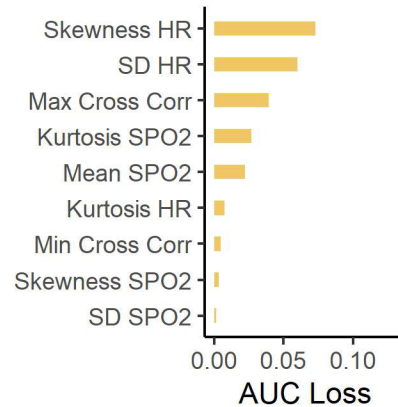
LR



NN



XG



RF

