


## SPECIAL ARTICLE



# AI models in clinical neonatology: a review of modeling approaches and a consensus proposal for standardized reporting of model performance

Ameena Husain<sup>1,12</sup>, Lindsey Knake<sup>2,12</sup>, Brynne Sullivan<sup>3,12</sup>, James Barry<sup>4</sup>, Kristyn Beam<sup>5</sup>, Emma Holmes<sup>6</sup>, Thomas Hooven<sup>7</sup>, Ryan McAdams<sup>8</sup>, Alvaro Moreira<sup>9</sup>, Wissam Shalish<sup>10</sup> and Zachary Vesoulis<sup>11</sup>

© The Author(s), under exclusive licence to the International Pediatric Research Foundation, Inc 2024

Artificial intelligence (AI) is a rapidly advancing area with growing clinical applications in healthcare. The neonatal intensive care unit (NICU) produces large amounts of multidimensional data allowing AI and machine learning (ML) new avenues to improve early diagnosis, enhance monitoring, and provide highly-targeted treatment approaches. In this article, we review recent clinical applications of AI to important neonatal problems, including sepsis, retinopathy of prematurity, bronchopulmonary dysplasia, and others. For each clinical area, we highlight a variety of ML models published in the literature and examine the future role they may play at the bedside. While the development of these models is rapidly expanding, a fundamental understanding of model selection, development, and performance evaluation is crucial for researchers and healthcare providers alike. As AI plays an increasing role in daily practice, understanding the implications of AI design and performance will enable more effective implementation. We provide a comprehensive explanation of the AI development process and recommendations for a standardized performance metric framework. Additionally, we address critical challenges, including model generalizability, ethical considerations, and the need for rigorous performance monitoring to avoid model drift. Finally, we outline future directions, emphasizing the importance of collaborative efforts and equitable access to AI innovations.

*Pediatric Research*; <https://doi.org/10.1038/s41390-024-03774-4>

## INTRODUCTION

With rapid improvements in the fields of data science and data analytics over the past decades, combined with ever-expanding accessible data in healthcare, artificial intelligence (AI) applications to help create solutions to clinical problems has become a new frontier in medical innovation.<sup>1,2</sup> To name a few, AI has been successfully used in drug discovery, robotic surgery, electronic health records (EHR), and more recently for outbreak prediction. There are 950 AI-enabled devices that have been approved by the Food and Drug Administration (FDA) as of August 2024, 16 common procedural terminology (CPT) codes can be used for reimbursement when using these devices, and over 100,000 articles on AI were published between 2022–2024.<sup>3,4</sup> Neonatology is ripe for the development of AI applications and is uniquely positioned to reap the benefits from AI's breakthroughs.<sup>5</sup> First, infants admitted to the neonatal intensive care unit (NICU), especially preterm infants, generate extensive, large-scale, and multidimensional datasets over the course of their hospitalization. This includes continuous physiologic data, imaging, laboratory,

and data from mechanical devices. Second, strong collaborative infrastructures already exist in neonatology, with many NICUs already sharing their data within well-established neonatal research networks. Third, prematurity is associated with a very high disease burden, in which morbidities are often complex entities that can be best understood by evaluating multi-dimensional data including biomedical signals, microbiome, and multi-omics in very large populations.

In the following review, we will provide a snapshot of some key clinical applications of AI in neonatology that address complex neonatal problems. This first section will highlight the wide range of AI models currently being employed in neonatal research and the highly variable measures of performance used to interpret and report the results. In the second and third sections, we will provide a systematic framework to guide the development and evaluation of AI models, including different measures of performance. In the final section, we highlight limitations and ongoing mitigation strategies to ensure safe and ethical implementation of clinical AI models in Neonatology.

<sup>1</sup>Division of Neonatology, Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, UT, USA. <sup>2</sup>Division of Neonatology, Department of Pediatrics, University of Iowa, Iowa City, IA, USA. <sup>3</sup>Division of Neonatology, Department of Pediatrics, University of Virginia School of Medicine, Charlottesville, VA, USA. <sup>4</sup>Division of Neonatology, Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO, USA. <sup>5</sup>Department of Neonatology, Beth Israel Deaconess Medical Center, Boston, MA, USA. <sup>6</sup>Division of Newborn Medicine, Department of Pediatrics, Mount Sinai Hospital, New York, NY, USA. <sup>7</sup>Division of Newborn Medicine, Department of Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. <sup>8</sup>Department of Pediatrics, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA. <sup>9</sup>Division of Neonatology, Department of Pediatrics, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA. <sup>10</sup>Division of Neonatology, Department of Pediatrics, Research Institute of the McGill University Health Center, Montreal Children's Hospital, Montreal, Canada. <sup>11</sup>Division of Newborn Medicine, Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, USA. <sup>12</sup>These authors contributed equally: Ameena Husain, Lindsey Knake, Brynne Sullivan. email: ameena.husain@hsc.utah.edu

Received: 30 October 2024 Accepted: 10 November 2024

Published online: 17 December 2024

## CLINICAL APPLICATIONS WITHIN NEONATOLOGY

### Neonatal sepsis

A promising application of AI in neonatology can be found in the development of early warning systems for late-onset neonatal sepsis, a condition that contributes significantly to preterm infant morbidity and mortality.<sup>6</sup> Neonatal sepsis challenges NICU clinicians because the signs and symptoms of neonatal infection are non-specific and frequently overlap with non-infectious conditions. Machine learning (ML) offers the advantage of using computers to detect subtle patterns in the pre-clinical phase of sepsis, as the systemic inflammatory response builds, but before illness becomes obvious and the need for intervention urgent.<sup>7</sup> Many sepsis risk prediction systems have been published, but few have been implemented or tested for generalizability using external validation. Researchers have used diverse approaches to modeling methods and data inputs.<sup>8–12</sup> Physiologic monitoring data are often considered most useful for early warning algorithms because they are generated by the patient and dynamically change as sepsis develops.<sup>13</sup> The autonomic response to systemic inflammation manifests as changes in cardiorespiratory control, including reduced heart rate variability and increased apnea with bradycardia and desaturation.<sup>14–16</sup> EHR data such as laboratory tests for biomarkers, physiologic support changes, and orders placed by providers can help with sepsis risk stratification and have good predictive performance for diagnosing late-onset sepsis.<sup>17</sup> However, EHR data are generated by provider action, requiring a threshold of patient change for activation, correlating better with severity of illness or prediction of outcome rather than providing early warning.

Heart rate characteristics monitoring is an example of ML applied to physiologic data for early warning of late-onset neonatal sepsis.<sup>18</sup> The Heart Rate Characteristics (HRC) Index was developed to predict the risk of sepsis diagnosis in the next 24 hours using logistic regression with algorithms to detect reduced heart rate variability and transient heart rate decelerations.<sup>19</sup> The model was commercialized as the HeRO System (MPSC, Charlottesville, VA) and received FDA 510 K clearance as a monitoring device. The HeRO trial tested the impact of implementing the HRC-based sepsis risk monitoring without threshold alerts or mandated actions. Infants were randomized to have the HRC index displayed or not displayed to the medical team from NICU admission until discharge, death, or 120 days postnatal age. The results of this 9-NICU randomized clinical trial (RCT) showed that HRC index display reduced all-cause mortality by 20%,<sup>20</sup> and death within 30 days of sepsis by 40%.<sup>21</sup> Despite these significant results of the only multicenter RCT to test an implemented ML system in the NICU, interpretation, acceptance, and adoption of this system is not widespread.<sup>22</sup> Thus, the future of AI for sepsis risk prediction depends on not only the algorithm but also pre- and post-implementation strategies and continuous evaluation of their impact on patient outcomes to achieve widespread adoption.

### Retinopathy of prematurity

Retinopathy of prematurity (ROP) is a serious eye disorder in premature infants that can lead to blindness if not detected and treated early. Traditional screening methods, which rely on experienced ophthalmologists performing bedside fundus photography or direct exam, face challenges due to the increasing number of at-risk infants and a shortage of qualified professionals, especially in non-tertiary care centers.<sup>23</sup> These limitations underscore the need for innovative approaches, such as the application of AI, to automate or even enhance ROP screening.

Deep neural networks (DNNs) are complex computational models that utilize multiple layers of interconnected nodes to learn and extract intricate patterns from data, enabling tasks such as image recognition and decision-making. To effectively train DNNs for ROP screening, diverse and extensive patient datasets

are needed to capture various ROP features and severities. Wang et al. developed DeepROP, an automated ROP detection system using two models called Id-Net for identification and Gr-Net for grading.<sup>24</sup> The DeepROP system showed promising results in a clinical setting, with Id-Net achieving a sensitivity of 84.91% and a specificity of 96.90% for ROP identification, while Gr-Net attained a sensitivity of 93.33% and a specificity of 73.63% for ROP grading. This system performed comparably to human ophthalmologists and supported telemedicine by facilitating hospital collaboration and pre-screening in non-specialized facilities.

Deep learning algorithms have the potential to enhance ROP prevention efforts worldwide, especially in resource-limited settings. This was demonstrated in a multinational validation study by Coyner et al. which evaluated an autonomous AI system using a deep learning algorithm for ROP screening.<sup>25</sup> The model, developed using data from the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) study, was validated on datasets from the United States and India. Notably, no infants were diagnosed with type 1 ROP before being flagged by the autonomous AI system. If fully implemented, this AI system could reduce physician workload associated with telemedicine examinations by up to 80%. However, widespread adoption will depend on investment in compatible digital cameras and further validation of the algorithm to ensure it can process images from future devices effectively.

Support Vector Machines (SVMs) have been effective in classifying plus disease, a severe indicator of ROP. These models identify the optimal hyperplane that separates different classes of data points based on their features. Ataer-Cansizoglu et al. used SVMs to achieve 95% accuracy in diagnosing plus and preplus disease by utilizing a novel feature representation based on Gaussian Mixture Models.<sup>26</sup> This approach enabled the differentiation between healthy and abnormal vascular features, showcasing their robustness in clinical applications. Despite their effectiveness, SVMs can become computationally intensive and less efficient with large datasets, as training time and memory requirements increase with the number of data points and features.

Researchers have also explored hybrid approaches to leverage multiple types of information for more accurate ROP diagnosis. For example, Hu et al. combined Convolutional Neural Networks (CNNs) with Long Short-Term Memory networks (LSTMs) to utilize both spatial and temporal information from infant eye exams, achieving 97% accuracy for ROP diagnosis.<sup>27</sup> This hybrid model integrates CNNs' strength in analyzing spatial features with LSTMs' ability to capture temporal changes, providing a comprehensive assessment of disease progression. Such approaches enhance diagnostic accuracy and reliability, particularly in complex cases where single-model approaches may fall short. Table 1 provides descriptions of these models for comparison.

The development of robust, generalizable ML models for ROP screening have been hampered by the use of small, single-center datasets and non-standardized imaging protocols and reference standards in studies. External validation on multi-center data is important for establishing the generalizability of these models, thus a standardized approach to image acquisition and the promotion of multicenter projects is needed for further development. A recent multicenter study validating the ROP.AI deep learning algorithm using over 8,000 retinal images from five Australian centers demonstrated the importance of external validation for assessing generalizability. ROP.AI achieved a sensitivity of 84% and a negative predictive value of 96% for detecting plus disease after optimizing the operating threshold. However, the study also highlighted challenges, such as reduced specificity (43%) and variability due to differences in image quality and grading practices across centers, emphasizing the need for further training to improve clinical applicability.<sup>28</sup> Additionally, affordable, compatible digital cameras are needed to address needs in low-resource settings.<sup>25</sup>

**Table 1.** Description of common artificial intelligence models.

Model (Acronym)	Description of Model
Logistic Regression (LR)	Classification model that models the log-odds of an outcome as the linear combination of one or more predictors
Decision Trees (DT)	Classification model that splits the data based on features, which creates branches to represent decision paths
Random Forest (RF)	Ensembles of multiple decision trees trained on different subsets of data with results aggregated
Gradient boosting machines (GBM)	Builds multiple trees sequentially with each tree correcting the errors of the previous tree
Support Vector Machines (SVMs)	Decision boundary (vector) that separates the categories (outcomes) and reduces errors
<b>Deep neural networks (DNNs)</b>	<b>Complex computational models that utilize multiple layers of interconnected nodes to learn and extract intricate patterns from data</b>
Convolutional Neural Network (CNN)	Type of DNN with Interconnected nodes (matrices of weights) based on the number of features in the model with multiple convolutional layers
Transformers	Type of DNN that use parallel processing to create an encoder to compress sequential data into an internal representation and then a decoder is used to regenerate the data
Recurrent Neural Network (RNN)	Type of CNN that can process sequential data by maintaining memory of previous inputs.
Long Short-Term Memory (LSTM)	Type of RNN with enhanced memory function for longer sequences

Continued research is needed to validate these tools prospectively and integrate them effectively into clinical workflows. With further refinement, ML-assisted ROP screening could expand access to expert-level diagnostics, especially in resource-limited settings.

### Necrotizing enterocolitis

Necrotizing enterocolitis (NEC) is a life-threatening intestinal disease that predominantly affects preterm infants.<sup>29</sup> As with sepsis, early clinical features of NEC can be subtle. With progression, which can be rapid or insidious, affected infants experience intestinal necrosis, hemodynamic instability, and multiorgan dysfunction requiring surgical resection of dead bowel. Long-term digestive, growth, and neurodevelopmental complications among survivors are common,<sup>30</sup> heightening the need for new techniques for NEC prevention, diagnosis, and treatment.

Multiple groups have used ML approaches to understand NEC pathogenesis and improve clinical management strategies. A challenge facing these efforts is technical heterogeneity between datasets and a concern that a clinical event labeled as “NEC” might represent different underlying pathologies that converge on a common acquired intestinal dysfunction and injury.<sup>31</sup> To develop a less biased labeling system, Gipson et al. performed unsupervised ML classification of 183 infants with acquired intestinal disease, identified based on imaging findings, ICD billing codes, or history of surgical abdominal intervention. This strategy yielded five disease subtype clusters, which the authors labeled low mortality, mature with inflammation, immature with high mortality, late injury at full feeds, and late injury with high rates of intestinal necrosis.<sup>32</sup> Future unbiased classification studies may succeed in refining the definition of NEC as a specific disease entity—precision that could improve future ML training efforts aimed at prediction and decision support.

In the absence of a new diagnostic paradigm, ML efforts so far have relied upon classical Bell’s staging of NEC to characterize disease severity. These phenotype definitions have been paired with different clinical and biological feature datasets in efforts to understand NEC pathogenesis, discover new biomarkers of developing or worsening disease, and undergird decision support tools. Some research groups have used ML algorithms to optimize features selected for inclusion in downstream ML prediction models.<sup>33</sup> Others have started with predetermined biological samples and specific analysis modalities—often using -omics approaches such as metabolomics or metagenomics—to generate large datasets for ML system training and testing.

Because basic research studies have linked perturbation of the intestinal microbiota to NEC,<sup>34</sup> several teams have trained ML systems on neonatal microbiota data. Using bacterial population structures as training and test features in a random forest model, then mining top taxonomic hits suggested by the model for metabolic signatures, Casaburi et al. identified elevated stool formate as a final common pathway of high-risk microbiota that predisposed to NEC.<sup>35</sup> Similarly, Olm et al. used a decision tree-based classifier (Table 1) to identify microbiota metagenomic features potentially linked to NEC, finding that bacterial replication rate and high abundance of specific *Klebsiella* spp. preceded disease onset.<sup>36</sup> Lin et al. applied a neural network instantiation of multiple instance learning to stool microbiota data from two previously reported cohorts of preterm infants, developing a model that—in simulations on test data—predicted NEC an average of eight days before disease onset.<sup>37</sup> Other groups have examined urine metabolomics,<sup>38</sup> stool metabolomics,<sup>39</sup> and ML training on a collection of abdominal radiographs as strategies to improve NEC diagnosis, prognostication, and decision support.<sup>40</sup> Near-term future goals for ML research on NEC include establishing broad generalizability of reliable prediction and diagnostic models. Beyond direct application of ML NEC models to reduce morbidity and mortality through faster diagnosis, interpretable NEC modeling might also allow new insights about disease pathogenesis by pinpointing molecular, clinical, or environmental factors that affect disease risk. A recently published review of NEC ML studies is recommended for interested readers.<sup>41</sup>

### Neurodevelopmental outcomes

Predicting brain injury and neurodevelopmental consequences are highly salient topics for providers and parents alike. While certain clinical factors (gestational age, hypoxia, inflammatory processes such as NEC and sepsis) are well known to increase the risk of adverse neurodevelopmental outcomes, they are imprecise predictors. Additionally, there is increasing recognition of neurodevelopmental impairment identified later in childhood despite absence of overt injury on imaging, suggesting that subtle and unidentified factors are at play. ML has been applied to several different aspects of neonatal neurocritical care, the most notable of which are seizure detection and electroencephalogram (EEG) background classification algorithms as well as deep learning MRI classifiers.

ML-based EEG tools represent perhaps the most successful application of AI to neonatal clinical care. Neonates experience more seizures than any other population group across the human lifespan, yet neonatal seizures are notoriously difficult to detect

without the use of EEG. While EEG represents the gold standard monitoring, it requires extensive physical and human capital and is not accessible in many settings. Limited channel amplitude integrated EEG (aEEG) is designed to bridge this gap, but unfortunately studies have demonstrated inadequate seizure detection capabilities. AI-based automated neonatal seizure detection algorithms have been available for nearly 25 years, including the Gotman and Navakatikyan algorithms, both of which have FDA approval and detect approximately 85% of neonatal seizures.<sup>42,43</sup> More recently, Boylan and colleagues have developed Algorithm for Neonatal Seizure Recognition (ANSeR) which uses a SVM classifier to identify seizures in multi-channel EEG recordings, detecting 96% of seizures with few false negatives.<sup>44</sup> In clinical trial, the algorithm allowed for far greater recognition of seizures compared to human-scoring alone.<sup>45</sup> Given the known relationship between increased seizure burden and adverse outcome, accurate and complete seizure recognition provides the best opportunity for timely treatment.

Challenges associated with the interpretation of imaging parallel those of EEG; systems require not just the operation of expensive equipment but also require a staff of experts able to provide consistent subjective interpretation. AI has rapidly found a place in the field of radiology; at least 5 different commercial stroke imaging platforms are available, most with FDA approval, which aid in localization of the lesion and identification of vessel occlusion to guide management.<sup>46</sup>

While there is a lack of similar commercialized tools in the neonatal space, many research efforts are underway leveraging ML to maximize the value of data collected by MRI. In one hybrid-AI study, deep learning was used in conjunction with detailed human-scored MRIs and clinical factors to identify the smallest number of features to predict motor outcomes, with putamen/globus pallidus injury, cord pH, and gestational age emerging as the most important predictive factors.<sup>47</sup> Lewis et al. used an MRI “template” to register MRIs of infants with hypoxic ischemic encephalopathy against an atlas of control infants.<sup>48</sup> Patterns of injury, termed “radiomic signatures,” were then identified region by region using quantitative comparison between cases and controls. This technique allows recognition of differences that are below the threshold of the human eye to detect or deviate from classical neuroradiology MRI reading methods. Tian et al. also utilized a CNN (Table 1) to identify “radiomics signatures” on atlas-registered T1 and T2 sequences in conjunction with one clinical factor (birth weight) to develop a novel risk prediction nomogram.<sup>49</sup> Finally, Lew et al. recently published a 4D CNN model which utilizes T1, T2, and diffusion sequences to predict death or neurodevelopmental impairment and achieves performance equal to or exceeding experienced neuroradiologists.<sup>50</sup>

Machine learning has the potential to revolutionize neonatal neurocritical care by improving the detection of subtle and overt brain injuries, offering more precise predictions of neurodevelopmental outcomes. These emerging technologies are poised to enhance diagnostic accuracy and expand the capabilities of neonatal neuroimaging and EEG analysis, driving a new era of data-driven care.

### Bronchopulmonary dysplasia

Bronchopulmonary dysplasia (BPD) is a chronic lung disease predominantly affecting very preterm infants, marked by disrupted alveolar and vascular development.<sup>51</sup> BPD arises from multiple contributing factors, including prolonged mechanical ventilation, oxidative stress, persistent inflammation, chorioamnionitis, suboptimal nutrition, and infection.<sup>52</sup> Despite significant advancements in neonatology, effective preventive treatments remain scarce.

AI is playing an increasingly important role in managing BPD, particularly in early detection and risk prediction. Advanced ML models are being trained to analyze diverse clinical data, to

identify infants most susceptible to BPD. An early example of this approach, developed using logistic regression approaches, is the BPD risk estimator ([https://neonatal.rti.org/index.cfm?fuseaction=BPD\\_Calculator2.start](https://neonatal.rti.org/index.cfm?fuseaction=BPD_Calculator2.start)); however, its utility is limited by relying on a handful of static variables, unlike daily estimators that leverage the extensive, dynamic data available in EHR systems.<sup>53</sup> For example, Montagna et al. demonstrated that AI, when combined with conventional statistics, identified key risk factors—such as low gestational age, need for mechanical ventilation, and abnormal umbilical artery flow—that are strongly predictive of BPD.<sup>54</sup> Similarly, a deep learning model by Chou et al. utilized chest radiographs to predict BPD in preterm infants with high accuracy, surpassing expert-level performance in early diagnosis.<sup>55</sup> Moreira et al. also showcased the integration of AI with gene expression data to create a transcriptomic signature capable of predicting BPD risk as early as five days after birth.<sup>56</sup> These predictive tools may empower clinicians to implement targeted preventive strategies and deliver more personalized interventions.

However, several challenges remain. One significant barrier, common in NICU research, is the variability in clinical data quality and availability, which can delay the generalizability of AI models across diverse populations and healthcare settings.<sup>57</sup> Moreover, the integration of AI into clinical practice requires robust validation through large-scale, multi-center collaborations to ensure the accuracy and safety of these predictive models.<sup>58</sup> Ethical concerns also emerge, particularly as they relate to BPD, given that sociodemographic factors such as race, ethnicity, sex, and socioeconomic status play a role in BPD risk and prognosis.<sup>59–62</sup> Additionally, while AI has demonstrated remarkable success in risk prediction and early diagnosis, there remains a gap in its application for guiding therapeutic interventions, particularly in developing personalized treatment plans for infants across different stages of BPD.<sup>63</sup> Addressing these challenges will require a collaborative effort from neonatologists, data scientists, and regulatory bodies to optimize AI's potential for improving outcomes in preterm infants at risk or with BPD.<sup>64</sup>

As research continues to confront these challenges, the collaboration between neonatology and AI holds promise for transforming BPD management. Leveraging AI-driven insights for personalized care strategies could reduce the incidence and severity of BPD, improve long-term pulmonary outcomes, and enhance the overall quality of life for premature neonates. Successful AI integration in BPD care will depend on ongoing interdisciplinary collaboration, rigorous validation, and a commitment to equity in algorithm development.

### Extubation readiness

Determining when to discontinue mechanical ventilation (MV) and extubate infants is challenging. A study in infants <1250 g found that 47% of patients were reintubated during hospitalization.<sup>65</sup> In infants born <27 weeks, extubation failure rates typically range anywhere from 30–60% but can reach as high as 80% in some centers.<sup>66</sup> Early extubation failure within the first postnatal week in extremely preterm infants is associated with significantly increased mortality rates (11–12%) compared to successful extubations (3%).<sup>66,67</sup> Failed extubations are also associated with higher risks of BPD, severe intracranial hemorrhage, longer hospitalizations, and longer duration of respiratory support.<sup>68–70</sup> Unfortunately, assessment of extubation readiness currently relies on clinical judgment, which is subjective and associated with significant practice variations. While some centers utilize spontaneous breathing trials (SBTs), the latter have not been shown to improve prediction of extubation success and often cause unintended clinical instability in the most immature patients.<sup>71</sup>

Numerous prediction models using logistic regression or ML methods have been developed to help clinicians determine the appropriate timing of extubation in preterm infants (Table 2).<sup>71–79</sup>



**Table 2.** Summary of previously published extubation success prediction models.

Year 1 <sup>st</sup> Author	Total (n)	Cohort Definition	EF rate	EF Definition	Type of Data	Best Performing Algorithm	Multi-center	CDS tool	External data	Best AUROC
2004 Meuller	183	BW 900 – 1500 g	20%	<48 h	LR	ANN	No	No	No	0.87
2012 Mikhno	179	GA 23 – 31 weeks	13%	<48 h	LR	MLR	No	No	No	0.87
2013 Meuller	486	BW 500 – 1500 g	12%	<48 h	LR	MLR	No	No	No	0.78
2018 Goel	66	GA < 35 weeks	27%	<72 h	HR	MLR	No	No	Yes	NR
2019 Gupta	312	BW ≤ 1250 g	27%	<120 h	LR	MLR	No	Yes	No	0.77
2020 Chakraborty	577	GA < 32 weeks	23%	<72 h	HR	MLR	Yes	Yes	Yes	0.72
2021 Cheng	186	GA 25 – 29 weeks	43%	<120 h	LR	MLR	Yes	No	Yes	0.82
2022 Dryer	177	BW ≤ 1250 g	32%	<120 h	LR	MLR	No	Yes	Yes	0.79
2022 Kanbar	241	GA < 28 weeks	18%	<72 h	HR	Random Forest Classifier	Yes	No	No	0.75
2022 Natarajan	1,348	BW < 2500 g	26%	<168 h	HR	XGBoost	No	No	No	0.82
2023 Chen	60	BW < 1500 g	22%	<72 h	LR	MLR	No	No	No	0.74
2023 Hoffman	89	GA < 28 weeks	26%	<72 h	HR	MLR	No	No	No	0.81
2023 Song	678	GA < 32 weeks	16%	<72 h	HR	MLR	No	No	Yes	0.89
2024 Brasher	110	GA < 30 weeks	24%	<168 h	HR	Random Forest Classifier	No	No	No	0.94

EF extubation failure, CDS clinical decision support, NR not reported, LR low-resolution, HR high-resolution, ANN Artificial Neural Network, MLR multivariable logistic regression, AUROC area under receiver operating characteristic, BW birthweight, GA gestational age, g grams, h hours

However, none of these models have been adopted routinely into clinical care, likely due to a number of limitations. One limitation to most of the studies is the short duration of time over which reintubation is tracked after extubation (48 to 168 hours, depending on the model). Longitudinal data shows that reintubations may occur up to day 14 after extubation due to respiratory-related causes in neonates.<sup>80</sup> Consequently, all of these models, to varying degrees, likely underestimate the true reintubation rate (i.e., the rate of extubation failure).

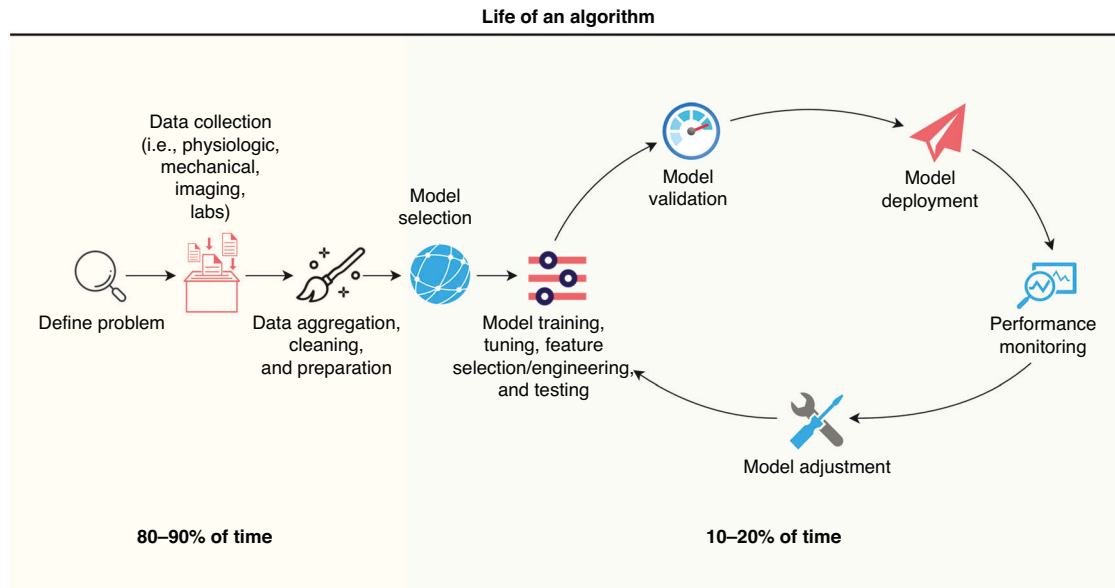
Another limitation is the number of infants included in the studies in Table 2 are low in the context of typical ML prediction model training. Additionally, the extubation failure rates mostly range between 12 – 26%; thus, there are even fewer failure cases for training the model. This imbalance may lead to overfitting of the model to the training data, especially when using methods that include many features as predictors. As expected, all of the models tested on external datasets had a drop in their area under the receiver operating characteristic (AUROC) curve when compared to their original model.<sup>73,77,80,81</sup> Although one study included >1000 infants in their model, this large sample size was achieved by including larger infants (who are less likely to fail extubation) and over a longer period, during which there may be significant changes in neonatal practices.<sup>74</sup> For those reasons, their model did not perform much better than using gestational age alone. The model with the highest accuracy, achieving an AUROC of 0.94 in a subgroup analysis, was underpowered as it only included 24 infants and was not validated on an external cohort. These models were likely overfit to the single center training data and thus, are unlikely to perform well at another site.<sup>82</sup>

Most studies in Table 2 used low-resolution clinical data (demographic information or at most, hourly ventilator settings and vital signs) to develop multivariable logistic regression models. Some studies have begun using high-resolution data such as continuous electrocardiography, thoracic impedance, pulse oximetry, and ventilator data showing evidence of improved sensitivity and specificity compared to using clinical variables alone.<sup>75</sup> Considering that measures of respiratory and heart rate variability in neonates have previously been associated with failed extubations,<sup>83,84</sup> inclusion of these continuous signals provides additive information about the infant's autonomic maturity and wellbeing as a surrogate of extubation readiness. However, inclusion of physiological data into prediction models remains technically and computationally challenging. In fact, most studies transformed continuous vital sign data into static values or descriptive statistics prior to incorporation into their models.<sup>74,76,79,81</sup>

Only a few studies have included multicenter data, and mostly in secondary data analyses.<sup>75,76,85</sup> Disappointingly, the few groups who have attempted external validation have also found that their AUROC decreases with the new data.<sup>76,81,85</sup> For example, Song et al. validated their data using a publicly available database and their AUROC decreased from 0.89 to 0.77.<sup>81</sup> Additionally, their poor negative predictive value indicates that 50% of babies predicted to fail extubation actually succeeded. Thus, simply basing performance on the AUROC of a model is flawed.

Two groups have created clinical decision support (CDS) tools based on their prediction models.<sup>72,76</sup> Neither of these models has been incorporated or implemented into routine clinical practice. Moreover, there have been no studies evaluating clinicians' opinions or acceptance of these CDS models, nor evaluating long-term clinical outcomes. The most recent example is the extubation.net prediction model developed by investigators at Wayne State University. When this model was formally validated in an external cohort from another center, the model performance decreased from an AUROC of 0.77 to 0.72.<sup>85</sup>

The future development of extubation failure prediction models should include large multi-center data with both clinical and high-



**Fig. 1** This diagram shows a high-level view of the process for artificial intelligence algorithm development. Of note, the initial steps of defining your problem, collecting data, and cleaning and preparing it for model development comprise the majority of the time during this process. After model deployment, continued monitoring and adjustments to the model will be required for ongoing effectiveness.

**Table 3.** Considerations of model selection based on type of data and application.

Data Type	Example data	Model	Applications
<b>Structured</b>	Patient demographics Tabular data Lab results	Machine learning models: Logistic regression Decision trees Random forests	Prediction Clinical decision support
<b>Complex</b>	Medical Images (CXR or MRI) Genomic data Complex physiologic signals	Deep learning models: Convolutional neural networks	Exploratory analysis Diagnosis
<b>Sequential</b>	Heart rate variability Respiratory rate over time Trends in vital sign data	Recurrent neural networks LSTM Transformers	Prediction Early warning systems

resolution continuous physiologic data incorporated into the models. Longitudinal studies will be needed to study and evaluate the everyday performance of an extubation readiness prediction model along with an implementation protocol, its broader clinical adoption, and its effect on neonatal outcomes.

## MODEL DEVELOPMENT

Model development in AI is similar to other statistical methods and follows a structured process (Fig. 1). For clinical model development, the aim is typically to predict specific clinical problems such as sepsis, ROP, or brain injury as described previously. A crucial step in the development of any model is the gathering and processing of data. As previously noted, a large sample size from diverse settings with a limited class imbalance in outcomes is optimal. However, clinical research data are messy, whether from clinical records, imaging, or physiological signals, and require careful review and preprocessing to ensure consistency in common values, correct alignment across patients and time, and to address missing values. This ensures standardized and consistent input for the model. Though laborious, the completeness of this step is the largest determination of the performance of any model.

### Model selection

Choosing the right model depends first and foremost on the nature of the data. Different types of models excel with different

datasets, so understanding the characteristics of the data is crucial. In general, we know that certain types of data work better for certain types of models (Table 3).

Choice of model will also depend upon the type of task being addressed. Generally, ML models are best at classification (i.e. assigning a yes/no or positive/negative). The more specialized the task, the more specialized the model will need to be to achieve it. A final consideration in model choice is degree of interpretability. There are differing opinion as to how explainable a model should be,<sup>86–88</sup> however, in clinical settings it is suggested to have some interpretability of the model so that when implemented, the model does not seem like such a black box.

An exhaustive discussion of available models is beyond the scope of this paper, but as shown in earlier examples, models can range from classification models using support vector machines in the case of ROP to convolutional neural networks to predict neurodevelopmental outcomes from MRIs. No one model will be perfect for a given research question, but each can be optimized with tools such as feature engineering and hyperparameter tuning. Feature engineering consists of selecting and potentially modifying the inputs to better suit the model, while hyperparameter tuning encompasses adjustments to characteristics of the model. If one considers the model to be a cake, then feature engineering is akin to adjusting the way you combine your ingredients and hyperparameter tuning is changes in the oven temperature. These adjustments should be made with careful consideration of the ramifications, and often require the support of a data scientist.

The final step in model selection is often a trial-and-error process, where multiple models are trained and compared based on performance metrics, discussed further below.

### Model construction

ML is a balancing act between detecting subtle patterns in data and overfitting (tuning the model to specific examples in the dataset with loss of performance on validation or external datasets). To achieve this, model construction consists of three stages: training, tuning, and testing.<sup>89</sup> During training, the model uses ground truth classifications to detect patterns which characterize those with and without the outcome and “learn” optimal parameters without learning patterns so specific that they are only found in the training data. Tuning allows the user to balance the model’s complexity with its ability to generalize well to new data. In this phase, adjustments are made to parameters and the slightly altered models are compared to one another. Finally testing is performed, during which the model is asked to classify data to determine the effectiveness of the model. Ideally, testing is performed on previously unseen data because re-using data from the training and tuning steps (also known as cross validation) may introduce bias in the model and overestimate performance.<sup>90,91</sup> While there may be instances when cross validation is mathematically preferable to designated testing data such as low sample size, there are many cases when use of cross validation may not be advisable.<sup>90</sup> We recommend the use of a completely withheld test set (ideally from a new source), and limiting resampling techniques, to the tuning process. For more complete discussion of optimal data splitting, see Xu and Goodacre.<sup>92</sup> In the absence of a true test set, statistics should be interpreted with care.

### MODEL PERFORMANCE EVALUATION

Effective evaluation of a new AI model and incorporation into routine clinical practice, hinge upon model performance evaluation, which in turn depends upon the use of statistical methods that may be unfamiliar to many clinicians. In this section, we break down the process of evaluating an AI algorithm. For the purposes of this review, we focus on the evaluation of binary classification models, the most common task for which ML is employed, in order to give a complete overview of the process and provide specific recommendations. While the principles of evaluation remain the same when applied to multiclass classification, regression, or other ML tasks, there are important differences in specific evaluation methods that are beyond the scope of this review. For a thorough discussion of the metrics employed for these tasks, see Ranio et al.<sup>93</sup>

Prior to examining evaluation metrics, we must ensure that the model is constructed in a way that will allow for interpretable performance metrics by taking into consideration the discussion of model construction above. After gaining confidence in our ability to assess model performance, we turn our attention to choosing performance metrics. First, one must choose to present either threshold dependent metrics (which require the researcher to decide on a threshold probability for distinguishing positive and negative cases), threshold independent metrics, or both. In the former case we use the following designations of a classification: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The most commonly used threshold dependent metrics utilize these designations to describe ratios of correctly classified instances within different populations as described in Table 4. Threshold independent metrics include AUROC or area under the precision-recall curve (AUPRC). Each of these metrics describes different aspects of the model and has its own advantages and pitfalls (Table 4). Care should be taken when interpreting recall, precision, and specificity, for example, as they reflect only a portion of the model’s performance. F1-score and

Matthews’ Correlation Coefficient (MCC) may provide a more general sense of performance, but present difficulty in translating to clinical implications.<sup>94</sup>

One major consideration in choosing metrics is the existence of class imbalance within the data set, which occurs whenever there is significant discrepancy between the number of positive and negative cases. Such imbalance is a regular occurrence in medical applications, particularly in the NICU where the majority of patients are affected by diseases related to development and conditions being monitored for are rare (typically less than 25% of the population). Accuracy and AUROC, in particular, suffer from significant overestimation in the presence of class imbalance.<sup>93,95</sup> Consider, for example, a disease with only 1% prevalence: a model that classifies all cases as negative would have 99% accuracy and an AUROC of 0.99, but would have no clinical utility. AUPRC is an alternative threshold independent metric that mathematically accounts for class imbalance, and thus provides an alternative metric from AUROC.

For authors, we recommend the following when deciding on which metrics to provide: 1) present multiple metrics when possible, particularly if presenting only threshold dependent metrics (consider all metrics in Tables 4), 2) consider the interpretability of the metric in conjunction with the aim of the model (for example, providing the recall for a model intended for screening may have greater clinical relevance than the MCC), 3) if using uncommon or complicated metrics, also provide common and widely understood metrics to avoid the appearance of hiding poor results behind complexity, and 4) take care not to overestimate performance when faced with significant class imbalance (for example, use of recall and precision rather than accuracy or AUPRC in conjunction with AUROC). Further, when choosing to present a threshold dependent metric, we recommend providing a rationale for the choice of threshold.

For readers, metric interpretation now follows simply from a close examination of the model construction and metric choice. Is the model constructed in such a way that the performance can be evaluated without bias? Are the presented metrics reasonable for the aim of the model? Do they provide complete information? Has class imbalance been considered? With these questions in mind the value of the chosen metric can be interpreted based on the guidance provided in Table 4.

### IMPORTANT CONSIDERATIONS WITH AI MODELS

A critical aspect of AI model deployment in healthcare is the need for continuous monitoring and maintenance to ensure algorithm effectiveness, as models are susceptible to “drift.” Model “drift” occurs when an AI model’s performance deteriorates over time due to shifts in clinical data, workflows, or the introduction of new technologies and interventions that were absent during the model’s initial training that change the clinical environment. In healthcare, even small changes, such as the adoption of new treatment protocols, new equipment, or diagnostic tools, can substantially reduce model accuracy. To mitigate the impact of “drift,” periodic monitoring and recalibration using up-to-date clinical data are essential to maintain model performance and ensure patient safety. Regularly updating the model in response to evolving clinical environments helps preserve its relevance and effectiveness.<sup>96</sup>

There are multiple ethical issues that need thoughtful consideration and evaluation when applying AI in pediatric healthcare including algorithmic bias, lack of a diverse study population in pediatric clinical trials, data privacy, transparency, fairness, equity, implicit and explicit racism.<sup>5,97,98</sup> Healthcare data, in particular, are often inherently biased, and if these biases are not adequately addressed or accounted for during model development, they risk being perpetuated and amplified by AI systems.

**Table 4.** Summary of Evaluation Performance Metrics

Threshold Dependent						
Metric	Equation*	Definition	Range	Interpretation	Pitfalls	Advantages
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Ratio of correctly classified instances to all instances	0-1	Overall correctness of the model	Highly subject to overestimation of performance in unbalanced data sets	Easily interpretable
Precision (aka positive predictive value (PPV))	$\frac{TP}{TP+FP}$	Ratio of correctly classified positive instances to all instances classified as positive	0-1	Likelihood that a positive classification is correct	Reflects only positive classification, so requires an additional metric to interpret fully	Easily interpretable
Negative predictive value (NPV)	$\frac{TN}{TN+FN}$	Ratio of correctly classified negative instances to all instances classified as negative	0-1	Likelihood that a negative classification is correct	Reflects only negative classification, so requires an additional metric to interpret fully	Easily interpretable
Recall (aka Sensitivity, true positive rate)	$\frac{TP}{TP+FN}$	Ratio of correctly classified positive instances to all positive instances	0-1	Ability to identify positives	Reflects only the detection of positives, so requires an additional metric to interpret fully	Easily interpretable
Specificity (aka true negative rate)	$\frac{TN}{TN+FP}$	Ratio of correctly classified negative instances to all negative instance	0-1	Ability to identify negatives	Reflects only the detection of negatives, so requires an additional metric to interpret fully	Easily interpretable
F1-Score	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision (positive) and recall	0-1	Overall model performance in correctly identifying positive cases	Impacted by class imbalance since the negative predictive value is not incorporated, though less so than accuracy	Improved performance in class imbalance
Matthews' correlation coefficient (MCC)	$\frac{(TN \times TP) - (FN \times FP)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	Correlation between the actual designations and assigned classifications	-1-1	Overall model performance, with 0 being no better than chance	Difficult to interpret in clinical setting	Well balanced even in cases of severe class imbalance
Threshold Independent						
Metric	X-axis	Y-axis	Range	Interpretation	Pitfalls	Advantages
Area under the receiver operating curve (AUROC)	False positive rate (1-specificity)	Recall	0.5-1	A value approaching 1 implies the model is highly successful at distinguishing positive and negative, while values approaching 0.5 imply the model is no better than random chance	Highly impacted by class imbalance. Cannot be used to interpret the performance of the model in negative cases.	Common, easily interpretable
Area under the precision recall curve (AUPRC)	Recall	Precision	Disease prevalence - 1	Values approaching 1 indicate the model is capable of distinguishing between true and false positives	Base level is the prevalence of disease, so cannot be compared between different diseases	Well balanced in cases of class imbalance

\*TN true negative, TP true positive, FN false negative, FP false positive.



Informed consent also presents a significant challenge, as AI becomes increasingly integrated into many aspects of healthcare such as medical device software found in MRI and CT scanners, predictive analytics and decision support, revenue cycle and supply chain operations management, patient and caregiver communication, and education. In some instances, health care data may move outside of a hospital or healthcare systems. Patients, caregivers, and guardians should be fully informed about the use of AI in their care, along with its potential risks and benefits. To foster trust and uphold ethical standards, a framework for notifying patients and obtaining informed consent regarding AI use is essential.<sup>99</sup> Incorporation of AI into standard consent for treatment represents the most straightforward approach, however it may become difficult or impossible for patients to “opt out” of AI, given the widespread and often invisible nature of AI in healthcare.

Recent updates on the regulation of AI in healthcare demonstrate a growing focus on ensuring safety, efficacy, and ethical considerations. Regulatory bodies have been increasingly proactive in publishing guidance documents and recognizing standards that demand rigorous development and validation of AI-enabled medical devices. This includes addressing risks such as bias and ensuring clinical validation in realistic settings. The European Union has introduced the AI Act, which adopts a risk-based approach to AI systems, complementing the General Data Protection Regulation to harmonize data protection and ethical AI principles.<sup>100</sup> The FDA continues to guide the regulatory approaches for AI in U.S. medicine focusing on the AI model's complete life cycle, safety, bias evaluation, model performance, risk, and user transparency.<sup>101</sup> Indeed, in 2021 the FDA released new guidance to manufacturers, broadly increasing the scope of computer software that must undergo review and clearance before marketing. The FDA, along with other government agencies are focusing on 4 main areas: safeguarding public health, developing regulatory approaches which support innovation, promoting the development of comprehensive guidelines, protocols, tools, and best practices, along with supporting research to evaluate and monitor AI performance. The Biden administration released an executive order that calls for the development of standards focused on ensuring that AI applications in healthcare maximize the benefits, while limiting or restricting risks.<sup>102</sup> A rigorous approach is needed if we are to promote and provide safe, ethical, and effective application of AI in healthcare.

## FUTURE DIRECTIONS

The future of AI in neonatal intensive care offers transformative advancements in patient care and outcomes, family engagement and experiences, and clinical workflows. AI-driven predictive analytics from continuous monitoring data will enable earlier detection of critical conditions like sepsis and necrotizing enterocolitis, providing real-time insights for more effective, prescriptive interventions. Multi-modal AI models will enhance diagnostic accuracy, personalized treatment plans, and tailored education for families and trainees using data from images, text, laboratory, visual, and physiological monitoring.

Integration of bedside monitoring with the electronic health records will reduce documentation burden, allowing clinicians and staff to focus on patient care. AI-augmented personalized medicine will deliver more precise, timely interventions based on individual physiologic, pathologic, genetic, and environmental factors. Additionally, AI will revolutionize medical education, shifting to adaptive learning models using tools like augmented reality, virtual reality, and AI tutors that allow for adaptive learning.

As AI models become more sophisticated, streamlining the approach to selecting and comparing them systematically with consistent evaluation metrics across studies will be key. This will

enable clinicians to choose the most suitable AI applications for specific clinical needs, facilitating better-informed decision-making. As AI matures, it will foster interdisciplinary collaboration, improve access to long-term neonatal follow-up care, and help reduce health disparities by providing equitable access to advanced technology. Ethical frameworks will be crucial in ensuring transparency, data privacy, and fairness, setting the stage for a new era in neonatal care.

By harnessing the transformative potential of AI, a new era of precision neonatology may emerge, where data-driven insights and tailored interventions come together to improve outcomes for our most vulnerable patients.

## REFERENCES

1. Md, H., Gs, C. & Kb, D. Three Epochs of Artificial Intelligence in Health Care. *JAMA*. 331. <https://doi.org/10.1001/jama.2023.25057>. (2024).
2. Haug, C. J. & Drazen, J. M. Artificial intelligence and machine learning in clinical medicine, 2023. *N. Engl. J. Med.* **388**, 1201–1208 (2023).
3. Joshi, G. et al. FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. *Electronics* **13**, 498 (2024).
4. Wu, K. et al. Characterizing the clinical adoption of medical AI devices through U.S. Insurance Claims. *NEJM AI* **1**, A0a2300030 (2023).
5. Sullivan, B. A. et al. Transforming neonatal care with artificial intelligence: challenges, ethical consideration, and opportunities. *J. Perinatol. J. Calif. Perinat. Assoc.* **44**, 1–11 (2024).
6. Flannery, D. D., Edwards, E. M., Coggins, S. A., Horbar, J. D. & Puopolo, K. M. Late-onset sepsis among very preterm infants. *Pediatrics* **150**, e2022058813 (2022).
7. Sullivan, B. A., Kausch, S. L. & Fairchild, K. D. Artificial and human intelligence for early identification of neonatal sepsis. *Pediatr. Res.* **93**, 350–356 (2023).
8. Cabrera-Quiros, L. et al. Prediction of late-onset sepsis in preterm infants using monitoring signals and machine learning. *Crit. Care Explor.* **3**, e0302 (2021).
9. Mani, S. et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *J. Am. Med. Inf. Assoc. JAMIA* **21**, 326–336 (2014).
10. Song, W. et al. A predictive model based on machine learning for the early detection of late-onset neonatal sepsis: development and observational study. *JMIR Med Inf.* **8**, e15965 (2020).
11. Garstman, A. G., Rodriguez Rivero, C. & Onland, W. Early detection of late onset sepsis in extremely preterm infants using machine learning: towards an early warning system. *Appl. Sci.* **13**, 9049 (2023).
12. Peng, Z. et al. DeepLOS: Deep learning for late-onset sepsis prediction in preterm infants using heart rate variability. *Smart Health* **26**, 100335 (2022).
13. Ba, S., Kd, F. Vital signs as physiometers of neonatal sepsis. *Pediatr. Res.* 91. <https://doi.org/10.1038/s41390-021-01709-x>. (2022).
14. Gholami, M. et al. Endotoxemia is associated with partial uncoupling of cardiac pacemaker from cholinergic neural control in rats. *Shock Augusta Ga.* **37**. <https://doi.org/10.1097/SHK.0b013e318240b4be>. (2012).
15. Herlenius, E. An inflammatory pathway to apnea and autonomic dysregulation. *Respir. Physiol. Neurobiol.* **178**, 449–457 (2011).
16. Fairchild, K. D., Srinivasan, V., Randall Moorman, J., Gaykema, R. P. A. & Goehler, L. E. Pathogen-induced heart rate changes associated with cholinergic nervous system activation. *Am. J. Physiol. - Regul. Integr. Comp. Physiol.* **300**, R330–R339 (2011).
17. Masino, A. J., et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PLoS One*. **14**. <https://doi.org/10.1371/journal.pone.0212665>. (2019).
18. Fairchild, K. D. & O'Shea, T. M. Heart rate characteristics: physiometers for detection of late-onset neonatal sepsis. *Clin. Perinatol.* **37**, 581–598 (2010).
19. Griffin, M. P. et al. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. *Pediatr. Res.* **53**, 920–926 (2003).
20. Moorman, J. R. et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *J. Pediatr.* **159**, 900–906.e1 (2011).
21. Fairchild, K. D. et al. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatr. Res.* **74**, 570–575 (2013).
22. Sullivan, B. A. & Keim-Malpass, J. BARRIERS to early detection of deterioration in hospitalized infants using predictive analytics. *Hosp. Pediatr.* **11**, e195–e198 (2021).
23. Barrero-Castillero, A., Corwin, B. K., VanderVeen, D. K. & Wang, J. C. Workforce shortage for retinopathy of prematurity care and emerging role of telehealth and artificial intelligence. *Pediatr. Clin. North Am.* **67**, 725–733 (2020).
24. Wang, J. et al. Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine* **35**, 361–368 (2018).

25. Coyner, A. S. et al. Multinational external validation of autonomous retinopathy of prematurity screening. *JAMA Ophthalmol.* **142**, 327–335 (2024).
26. Ataer-Cansizoglu, E. et al. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the “i-ROP” System and image features associated with expert diagnosis. *Transl. Vis. Sci. Technol.* **4**, 5 (2015).
27. Hu, J., Chen, Y., Zhong, J., Ju, R. & Yi, Z. Automated analysis for retinopathy of prematurity by deep neural networks. *IEEE Trans. Med Imaging* **38**, 269–279 (2019).
28. Bai, A. et al. Multicenter validation of deep learning algorithm ROP.AI for the automated diagnosis of plus disease in ROP. *Transl. Vis. Sci. Technol.* **12**, 13 (2023).
29. Neu, J. & Walker, W. A. Necrotizing enterocolitis. *N. Engl. J. Med* **364**, 255–264 (2011).
30. Han, S. M. et al. Long-term outcomes of severe surgical necrotizing enterocolitis. *J. Pediatr. Surg.* **55**, 848–851 (2020).
31. Neu J. Introduction and historical aspects and where may we be going in the future: getting rid of necrotizing enterocolitis. *Pediatr. Med.* **7**. <https://doi.org/10.21037/pm-23-30>. (2024).
32. Gipson, D. R. et al. Reassessing acquired neonatal intestinal diseases using unsupervised machine learning. *Pediatr. Res.* **96**, 165–171 (2024).
33. Song, J. et al. Framework for feature selection of predicting the diagnosis and prognosis of necrotizing enterocolitis. *PLoS ONE* **17**, e0273383 (2022).
34. Pammi, M. et al. Intestinal dysbiosis in preterm infants preceding necrotizing enterocolitis: a systematic review and meta-analysis. *Microbiome* **5**, 31 (2017).
35. Casaburi, G. et al. Metabolic model of necrotizing enterocolitis in the premature newborn gut resulting from enteric dysbiosis. *Front. Pediatr.* **10**, 893059 (2022).
36. Olm, M. R. et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria. *Sci. Adv.* **5**, eaax5727 (2019).
37. Lin, Y. C., Salleb-Aouissi, A. & Hooven, T. A. Interpretable prediction of necrotizing enterocolitis from machine learning analysis of premature infant stool microbiota. *BMC Bioinforma.* **23**, 104 (2022).
38. Sylvester, K. G. & Moss, R. L. Urine biomarkers for necrotizing enterocolitis. *Pediatr. Surg. Int* **31**, 421–429 (2015).
39. Rusconi, B. et al. Gut Sphingolipid Composition as a Prelude to Necrotizing Enterocolitis. *Sci. Rep.* **8**, 10984 (2018).
40. Gao W., et al. Multimodal AI System for the Rapid Diagnosis and Surgical Prediction of Necrotizing Enterocolitis. *IEEE Access*. PP:1-1. <https://doi.org/10.1109/ACCESS.2021.3069191>. (2021).
41. McElroy, S. J. & Lueschow, S. R. State of the art review on machine learning and artificial intelligence in the study of neonatal necrotizing enterocolitis. *Front. Pediatr.* **11**, 1182597 (2023).
42. Gotman, J., Flanagan, D., Rosenblatt, B., Bye, A. & Mizrahi, E. M. Evaluation of an automatic seizure detection method for the newborn EEG. *Electroencephalogr. Clin. Neurophysiol.* **103**, 363–369 (1997).
43. Navakatikyan, M. A. et al. Seizure detection algorithm for neonates based on wave-sequence analysis. *Clin. Neurophysiol.* **117**, 1190–1203 (2006).
44. Temko, A., Thomas, E., Marnane, W., Lightbody, G. & Boylan, G. EEG-based neonatal seizure detection with Support Vector Machines. *Clin. Neurophysiol.* **122**, 464–473 (2011).
45. Pavel, A. M. et al. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. *Lancet Child Adolesc. Health* **4**, 740–749 (2020).
46. Soun, J. E. et al. Artificial intelligence and acute stroke imaging. *Am. J. Neuroradiol.* **42**, 2–11 (2021).
47. Vesoulis, Z. A. et al. Deep learning to optimize magnetic resonance imaging prediction of motor outcomes after hypoxic-ischemic encephalopathy. *Pediatr. Neurol.* **149**, 26–31 (2023).
48. Lewis J., et al. Automated Neuroprognostication via Machine Learning in Neonates with Hypoxic-Ischemic Encephalopathy. <https://doi.org/10.1101/2024.05.07.24306996>. (2024).
49. Tian, T. et al. Graphic intelligent diagnosis of hypoxic-ischemic encephalopathy using mri-based deep learning model. *Neonatology* **120**, 441–449 (2023).
50. Lew, C. O. et al. Artificial intelligence outcome prediction in neonates with encephalopathy (AI-OPiNE). *Radio. Artif. Intell.* **6**, e240076 (2024).
51. Thébaud, B. et al. Bronchopulmonary dysplasia. *Nat. Rev. Dis. Prim.* **5**, 78 (2019).
52. Dini, G., Ceccarelli, S. & Celi, F. Strategies for the prevention of bronchopulmonary dysplasia. *Front. Pediatr.* **12**, 1439265 (2024).
53. Greenberg R. G., et al. Online clinical tool to estimate risk of bronchopulmonary dysplasia in extremely preterm infants. *Arch. Dis. Child Fetal Neonatal Ed.* Published online June 21, fetalneonatal-2021-323573. <https://doi.org/10.1136/archdischild-2021-323573>. (2022).
54. Montagna S., et al. Combining artificial intelligence and conventional statistics to predict bronchopulmonary dysplasia in very preterm infants using routinely collected clinical variables. *Pediatr. Pulmonol.* Published online August 16, <https://doi.org/10.1002/ppul.27216>. (2024).
55. Chou, H. Y. et al. Deep Learning Model for Prediction of Bronchopulmonary Dysplasia in Preterm Infants Using Chest Radiographs. *J. Imaging Inform. Med.* Published online March 18, <https://doi.org/10.1007/s10278-024-01050-9>. (2024).
56. Moreira, A. et al. Development of a peripheral blood transcriptomic gene signature to predict bronchopulmonary dysplasia. *Am. J. Physiol. Lung Cell Mol. Physiol.* **324**, L76–L87 (2023).
57. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* **17**, 195 (2019).
58. Schwabe, D., Becker, K., Seyferth, M., Klaub, A. & Schaeffter, T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *NPJ Digit. Med.* **7**, 203 (2024).
59. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Patterns N Y N. 2*. <https://doi.org/10.1016/j.patter.2021.100347>. (2021).
60. Collaco, J. M. et al. Socio-economic factors and outcomes in chronic lung disease of prematurity. *Pediatr. Pulmonol.* **46**, 709–716 (2011).
61. Patel, P., Ellefson, A. & Paul, D. A. Racial Disparities Among Predicted Bronchopulmonary Dysplasia Risk Outcomes in Premature Infants Born <30 Weeks Gestation. *Health Equity* **7**, 825–830 (2023).
62. Deschamps, J. et al. Neighborhood Disadvantage and Early Respiratory Outcomes in Very Preterm Infants with Bronchopulmonary Dysplasia. *J. Pediatr.* **237**, 177–182.e1 (2021).
63. Schork N. J. Artificial Intelligence and Personalized Medicine. In: Von Hoff D. D., Han H., eds. *Precision Medicine in Cancer Therapy*. Springer International Publishing; 265-283. [https://doi.org/10.1007/978-3-030-16391-4\\_11](https://doi.org/10.1007/978-3-030-16391-4_11). (2019).
64. Askin, S., Burkhalter, D., Calado, G. & El Dakrouni, S. Artificial Intelligence Applied to clinical trials: opportunities and challenges. *Health Technol.* **13**, 203–213 (2023).
65. Shalish, W. et al. Patterns of reintubation in extremely preterm infants: a longitudinal cohort study. *Pediatr. Res.* **83**, 969–975 (2018).
66. Shalish, W. et al. Age at First Extubation Attempt and Death or Respiratory Morbidities in Extremely Preterm Infants. *J. Pediatr.* **252**, 124–130.e3 (2023).
67. Berger, J., Mehta, P., Bucholz, E., Dziura, J. & Bhandari, V. Impact of early extubation and reintubation on the incidence of bronchopulmonary dysplasia in neonates. *Am. J. Perinatol.* **31**, 1063–1072 (2014).
68. Chawla, S. et al. Markers of Successful Extubation in Extremely Preterm Infants, and Morbidity After Failed Extubation. *J. Pediatr.* **189**, 113–119.e2 (2017).
69. Manley, B. J., Doyle, L. W., Owen, L. S. & Davis, P. G. Extubating Extremely Preterm Infants: Predictors of Success and Outcomes following Failure. *J. Pediatr.* **173**, 45–49 (2016).
70. Epstein, S. K., Ciubotaru, R. L. & Wong, J. B. Effect of failed extubation on the outcome of mechanical ventilation. *Chest* **112**, 186–192 (1997).
71. Shalish, W. et al. Assessment of Extubation Readiness Using Spontaneous Breathing Trials in Extremely Preterm Neonates. *JAMA Pediatr.* **174**, 178–185 (2020).
72. Gupta, D. et al. A predictive model for extubation readiness in extremely preterm infants. *J. Perinatol.* **39**, 1663–1669 (2019).
73. Mueller M., Wagner C. C., Stanislaus R., Almeida J. S. Machine learning to predict extubation outcome in premature infants. *Proc Int Jt Conf Neural Netw Co-Spons Jpn Neural Netw Soc JNNS AI Int Jt Conf Neural Netw.* 2013:1. <https://doi.org/10.1109/IJCNN.2013.6707058>. (2013).
74. Natarajan, A. et al. Prediction of extubation failure among low birthweight neonates using machine learning. *J. Perinatol. J. Calif. Perinat. Assoc.* **43**, 209–214 (2023).
75. Kanbar, L. J. et al. Automated prediction of extubation success in extremely preterm infants: the APEX multicenter study. *Pediatr. Res.* **93**, 1041–1049 (2023).
76. Chakraborty, M., Watkins, W. J., Tansey, K., King, W. E. & Banerjee, S. Predicting extubation outcomes using the Heart Rate Characteristics index in preterm infants: a cohort study. *Eur. Respir. J.* **56**, 1901755 (2020).
77. Mueller, M. et al. Predicting extubation outcome in preterm newborns: a comparison of neural networks with clinical expertise and statistical modeling. *Pediatr. Res.* **56**, 11–18 (2004).
78. Mikhno, A. & Ennett, C. M. Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf.* 2012;2012:5094-5097.
79. Hoffman, S. B. et al. Autonomic markers of extubation readiness in premature infants. *Pediatr. Res.* **93**, 911–917 (2023).
80. Mueller, M., Almeida, J. S., Stanislaus, R. & Wagner, C. L. Can machine learning methods predict extubation outcome in premature infants as well as clinicians? *J. Neonatal Biol.* **2**, 1000118 (2013).
81. Song, W. et al. Development and validation of a prediction model for evaluating extubation readiness in preterm infants. *Int J. Med Inf.* **178**, 105192 (2023).
82. Brasher, M. et al. Predicting extubation readiness in preterm infants utilizing machine learning: a diagnostic utility study. *J. Pediatr.* **271**, 114043 (2024).

83. Silva, M. G. F., Gregório, M. L. & de Godoy, M. F. Does heart rate variability improve prediction of failed extubation in preterm infants? *J. Perinat. Med* **47**, 252–257 (2019).
84. Goel, N., Chakraborty, M., Watkins, W. J. & Banerjee, S. Predicting extubation outcomes-a model incorporating heart rate characteristics index. *J. Pediatr.* **195**, 53–58.e1 (2018).
85. Dryer, R. A. et al. Evaluation and validation of a prediction model for extubation success in very preterm infants. *J. Perinatol. J. Calif. Perinat. Assoc.* **42**, 1674–1679 (2022).
86. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* **3**, e745–e750 (2021).
87. Chaddad, A., Peng, J., Xu, J. & Bouridane, A. Survey of Explainable AI Techniques in Healthcare. *Sensors* **23**, 634 (2023).
88. Bienefeld, N. et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. *NPJ Digit Med.* **6**, 94 (2023).
89. Hicks, S. A. et al. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **12**, 5979 (2022).
90. Kuhn M., Johnson K. *Applied Predictive Modeling*. Springer; <https://doi.org/10.1007/978-1-4614-6849-3>. (2013).
91. Westerhuis, J. et al. Assessment of PLS-DA cross validation. *Metabolomics* **4**, 81–89 (2008).
92. Xu, Y. & Goodacre, R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J. Anal. Test.* **2**, 249–262 (2018).
93. Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **14**, 6086 (2024).
94. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
95. Ozenne, B., Subtil, F. & Maucourt-Boulch, D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.* **68**, 855–859 (2015).
96. M S, A. R., Nirmala, C. R., Aljohani, M. & Sreenivasa, B. R. A novel technique for detecting sudden concept drift in healthcare data using multi-linear artificial intelligence techniques. *Front Artif. Intell.* **5**, 950659 (2022).
97. Brewster, R. C. L. et al. Race and ethnicity reporting and representation in pediatric clinical trials. *Pediatrics* **151**, e2022058552 (2023).
98. van Genderen, M. E. et al. Charting a new course in healthcare: early-stage AI algorithm registration to enhance trust and transparency. *NPJ Digit Med* **7**, 119 (2024).
99. Rose, S. L. & Shapiro, D. An ethically supported framework for determining patient notification and informed consent practices when using artificial intelligence in health care. *Chest* **166**, 572–578 (2024).
100. Meszaros, J., Minari, J. & Huys, I. The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. *Front Genet* **13**, 927721 (2022).
101. FDA. Artificial Intelligence & Medical Products: How CBER, CDER, CDRH, and OCP are Working Together. Published online March 2024. <https://www.fda.gov/media/177030/download>
102. Mello, M. M., Shah, N. H. & Char, D. S. President Biden's executive order on artificial intelligence-implications for health care organizations. *JAMA* **331**, 17–18 (2024).

## AUTHOR CONTRIBUTIONS

All listed authors made substantial contributions to conception and design of the manuscript and acquisition of information presented. All authors contributed to drafting the article, revising it critically for important intellectual content, and final approval of the version to be published.

## FUNDING

Alvaro Moreira - National Institutes of Health (NIH) Eunice Kennedy Shriver National Institute of Child Health and Human Development; Award number: K23 HD101701.  
Brynne Sullivan - National Institutes of Health (NIH) Eunice Kennedy Shriver National Institute of Child Health and Human Development; Award number: K23 HD097254.

## COMPETING INTERESTS

The authors declare no competing interests.

## PATIENT CONSENT

No patient consent was required.

## ADDITIONAL INFORMATION

**Correspondence** and requests for materials should be addressed to Ameena Husain.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.