

Review

Validation of Clinical Prediction Models with Repeated-Measures Predictors: A Methodological Framework for Neonatal Digital Twins in IoT Environments

Joseph L. Hagan 

Department of Pediatrics, Section of Neonatology, Baylor College of Medicine, Houston, TX 77030, USA; jlhagan@bcm.edu

Featured Application

Selected components of the validation pipeline described in this review are being implemented in the open-source R package *cawCV*, which provides researchers with cluster-aware cross-validation tools for prediction models built on repeated-measures predictor data.

Abstract

Background: The neonatal intensive care unit (NICU) is one of medicine's most data-rich Internet of Things (IoT) environments, and monitoring digital twins (DTs) that predict clinical events from continuously acquired physiologic streams are a leading proposed application. Methodological standards for validating the prediction models that underlie such DTs remain poorly defined. A monitoring DT can be no more trustworthy than the validation of the prediction models it contains. NICU physiologic data are characterized by repeated within-subject measurements and clustering by infant and center, structural features that violate the independence assumptions of standard prediction model validation methods. **Methods:** This methodological review examines validation strategies for clinical prediction models with repeated-measures predictors, with emphasis on cross-validation approaches that preserve validity under within-subject dependence, calibration-aware model evaluation, and feature selection stability. **Results:** Several methodological gaps are identified across the monitoring DT development pipeline. Recent simulation evidence quantifying the optimism produced by naive cross-validation in this setting is summarized and positioned as the foundation for the validation strategies recommended here. **Conclusions:** The NICU DT Validation Framework (NICU-DTVF) is proposed, articulating the validation conditions a trustworthy monitoring DT should satisfy, at a level of methodological maturity that is stated explicitly for each component. An open-source R package (*cawCV*) implementing cluster-aware validation as one component of that framework is under development. The validation principles described are general and extend to clustered, repeated-measures prediction problems beyond the neonatal setting.



Academic Editor: Samuel Okegbile

Received: 22 June 2026

Revised: 13 July 2026

Accepted: 15 July 2026

Published: 17 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: digital twin; monitoring digital twin; neonatal intensive care; clinical prediction model; cross-validation; repeated measures; cluster-aware validation; calibration; Internet of Things; machine learning

1. Introduction

The neonatal intensive care unit (NICU) is one of the most densely instrumented IoT environments in clinical medicine, with each critically ill infant continuously monitored by bedside devices generating heterogeneous, high-frequency physiologic data streams that present the same challenges of real-time acquisition, integration, and predictive modeling that define IoT systems in industrial and engineering domains. These signals range from high-frequency waveform data such as ECG and plethysmography to lower-frequency device outputs and discrete electronic health record events including laboratory measurements, medication administrations, and nursing assessments. Over a hospitalization lasting weeks to months, these multimodal data sources collectively generate millions of observations per patient, making the NICU a highly instrumented clinical environment analogous to industrial IoT systems where DT architectures were first developed.

A DT, in its most general formulation, is a computational model that mirrors a physical entity in sufficient fidelity to facilitate simulation, prediction, and decision support [1]. In the manufacturing and engineering literature, DT frameworks are mature: architectures for data acquisition, model calibration, real-time synchronization, and deployment have been systematically described [2,3]. The translation of these frameworks to clinical medicine has accelerated considerably in recent years, with proposed applications ranging from individualized drug dosing to *in silico* surgical planning [4,5]. In neonatal medicine specifically, Pammi and colleagues [6] articulated a compelling vision in which digital twins and synthetic patient data could empower pediatric clinical trials by providing virtual control arms, reducing enrollment burdens, and enabling more rapid evaluation of interventions for one of medicine's most vulnerable populations. Realizing this vision depends critically on the prediction modeling infrastructure that this review addresses. The monitoring DT application, in which a computational model synchronizes continuously with NICU IoT data streams to produce updated risk predictions, is a natural application of the IoT environment the NICU already represents and the focus of this review. Following the taxonomy of Drummond and Gonsard [7], DTs may be classified as simulated or monitoring variants; the monitoring DT, which performs factual prediction from continuously updated physiologic streams, is the sole focus here, and the simulated DT, which performs counterfactual prediction under hypothetical interventions, is outside the scope of this review for reasons detailed in Section 2.3. This scope limitation is reflected in Figure 1.

That vision is scientifically well-motivated. Adverse outcomes among extremely preterm infants are frequent, clinically consequential, and often preceded by physiologic changes detectable in continuously monitored data before they are clinically apparent, as has been demonstrated for late-onset sepsis [8]. What remains absent, however, is the validated methodological scaffolding needed to build such a monitoring DT responsibly. The Pammi et al. viewpoint [6], like much of the DT clinical literature, focuses appropriately on the conceptual case and the ethical prerequisites. It does not, and was not intended to, specify the supervised learning strategies best suited to high-dimensional neonatal IoT data, or the cross-validation approaches that preserve validity when data are clustered by infant and center and temporally autocorrelated within infants. These gaps are not peripheral implementation details; they are the critical unsolved problems that separate the DT vision from realizable clinical application. This review argues that the primary barrier to neonatal digital twin implementation is not data availability but the absence of validated methodological standards.

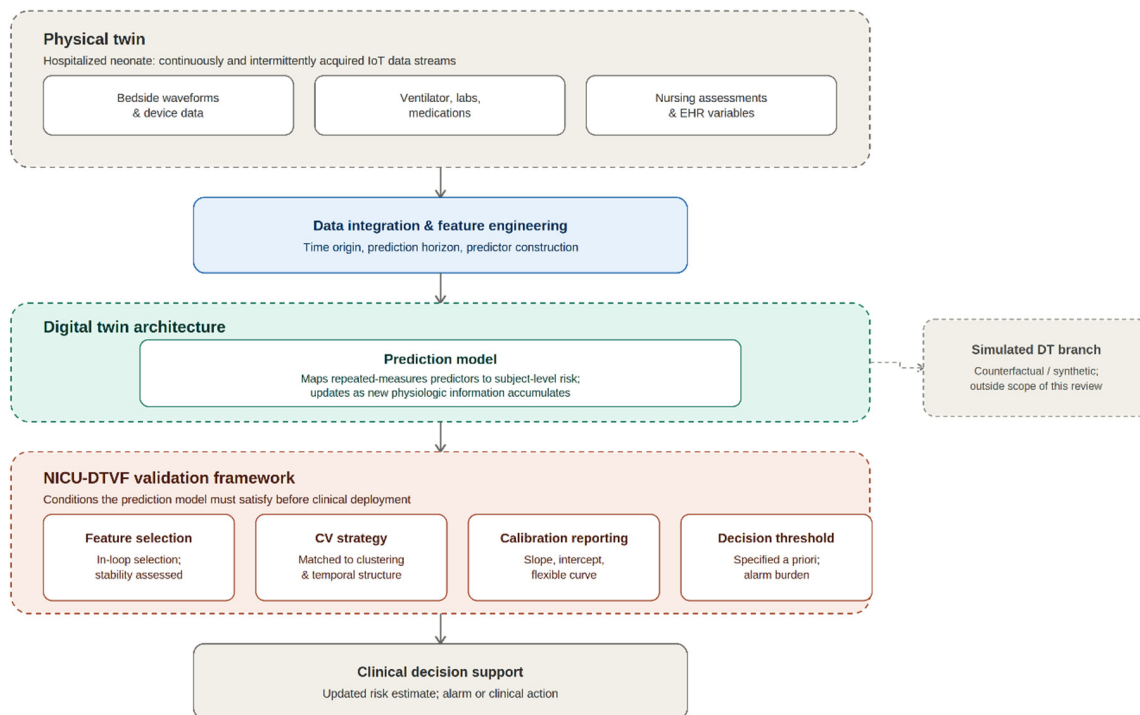


Figure 1. Conceptual NICU-DTVF architecture for an IoT-enabled neonatal monitoring digital twin. The pipeline begins with the physical twin, the hospitalized neonate, whose physiologic state is represented by continuously and intermittently acquired NICU data streams, including bedside waveform and device data, ventilator outputs, laboratory results, medication records, nursing assessments, and EHR-derived clinical variables. These heterogeneous IoT data streams enter a data-integration and feature-engineering layer, where time origin, prediction horizon, preprocessing, and derived repeated-measures predictors must be defined before model development. The machine-learning prediction model maps these time-anchored repeated-measures predictors to a subject-level clinical outcome and forms the predictive core of the monitoring digital twin, which updates risk estimates as new physiologic information accumulates. The NICU DT Validation Framework (NICU-DTVF) is shown as a set of validation gates embedded within this pipeline: feature selection performed within the resampling loop and assessed for stability; cluster-aware and time-aware validation matched to the infant-, center-, and temporal-dependence structure of the data; calibration-inclusive performance reporting; and a priori decision-threshold and alarm-burden evaluation before clinical deployment. Cross-cutting architecture and governance requirements, including privacy protection, interoperability, edge or federated learning options, regulatory monitoring, and prospective workflow evaluation, apply across the pipeline. The dashed simulated-DT branch indicates that counterfactual or synthetic-data applications represent a distinct simulated digital twin use case and require additional causal and fidelity-validation methods outside the primary scope of this review.

This problem has structural parallels to challenges well established in other high-dimensional biomedical settings. In genomic classification, the combination of large feature spaces, small effective sample sizes, and non-independent observations created a methodological crisis that took over a decade to partially resolve through the development of penalized regression, ensemble methods, and rigorous cross-validation protocols [9–11]. The NICU DT problem shares this structure: feature dimensionality is high, effective sample sizes for rare clinical outcomes are small, observations within infants are serially correlated, and the consequences of methodological shortcuts (e.g., overfitted models, optimistically biased performance estimates, poorly calibrated predictions) are clinically consequential [12]. The lessons learned from high-dimensional genomic prediction are directly applicable here, yet their translation to the DT-IoT context has not been systematically described.

Validation before deployment is, moreover, a recognized general requirement for DTs across engineering domains, from pre-execution verification in robot control systems [13] to reliability evaluation in condition-monitoring DTs for power electronics [14], a gap the National Academies have identified as foundational across domains [15]. The methodological contribution of this review is to specify what that validation requirement demands in the neonatal monitoring DT context, where the data structure, repeated within-subject physiologic measurements clustered by infant and center, renders standard validation methods invalid.

Standard clinical prediction modeling methodology, including penalized regression for cross-sectional outcome prediction, temporal external validation, and nomogram-based deployment, is well established for single-observation-per-subject clinical outcomes [16,17]. The methodological gap addressed by this review concerns the next step: validation of prediction models constructed from repeated within-subject measurements of physiologic data, for which the independence assumptions underlying standard validation methods are violated. Recent simulation work by the present author has quantified the optimism produced when these assumptions are ignored, providing the foundation on which the cross-validation recommendations developed in this review are constructed [18]. The machine learning methodology needed to construct and validate neonatal DTs in IoT-enabled NICU environments is examined here, with particular attention to three problem areas where the gap between current DT-IoT practice and neonatal clinical requirements is most acute: cross-validation strategies for clustered, temporally autocorrelated data; feature selection under small effective sample sizes; and calibration-aware model evaluation. Following the taxonomy of Drummond and Gonsard [7], DTs may be classified as simulated or monitoring variants; this review focuses on the monitoring DT, which performs factual prediction of clinical events from continuously updated physiologic streams and is the variant for which the validation methodology described here is directly applicable. The simulated DT, which performs counterfactual prediction under hypothetical interventions, raises a distinct set of identification and validation problems that fall outside the scope of this review and are noted where relevant. A structured framework integrating clinical prediction model standards [16,19] with DT-IoT methodology is proposed, concluded with a prioritized research agenda. The aim is not to survey what has been done, since the existing DT clinical literature, including comprehensive surveys of DT applications, methods, and challenges spanning cardiovascular, neurological, oncological, and other physiological systems, has been reviewed elsewhere [4,20,21], but to specify what must be done before neonatal DTs can be built responsibly and deployed credibly. Although the examples throughout are drawn from neonatal intensive care, the validation principles described are general: they apply to any clinical prediction model built on repeated-measures or clustered data, including adult intensive care, electronic health record-derived cohorts, and wearable-device monitoring. The conceptual scope of this review, encompassing the full neonatal DT pipeline from the physical twin to clinical application, is summarized in Figure 1, with the methodological problem areas detailed in subsequent sections corresponding to the ML pipeline stage.

2. The NICU as a DT-IoT Use Case: Data Landscape and Clinical Requirements

2.1. The Physical Twin Substrate: IoT Data Streams in the NICU

Understanding what data a neonatal DT must model begins with a clear accounting of what the physical twin, the hospitalized preterm infant, actually generates. The NICU physiologic data environment is characterized by extreme heterogeneity across at least three dimensions: sampling rate, data type, and clinical signal density (i.e., the amount of

clinically meaningful information contained in the data per unit time). High-frequency continuous streams include electrocardiographic waveforms, pulse oximetry plethysmography, and invasive arterial pressure traces, which may be sampled at 125–500 Hz and produce tens of millions of data points per patient per day. At intermediate frequencies, respiratory monitoring systems capture ventilator-derived parameters (tidal volume, peak inspiratory pressure, compliance, resistance) at breath-by-breath resolution, typically on the order of tens of observations per minute. At the low-frequency extreme, discrete clinical events (e.g., laboratory draws, medication administrations, nursing assessments, radiographic interpretations) are recorded as structured or semi-structured entries in the electronic health record (EHR), often with irregular and clinically driven timing rather than fixed intervals.

These streams are not independent. Physiologic parameters in the NICU are coupled in ways that reflect underlying pathophysiology: heart rate variability patterns anticipate late-onset sepsis by 12–24 h [8]; cerebral oxygenation trends co-vary with mean arterial pressure in the setting of impaired autoregulation [22]; and ventilator-derived compliance trajectories track surfactant response and predict extubation readiness [23]. A neonatal DT that models each stream in isolation fails to capture this clinical richness. The coupled, serially correlated structure of these data is precisely what makes their validation non-trivial, an observation that bears directly on the cross-validation framework developed in Section 4.

2.2. Clinical Outcomes of Interest and Their Statistical Implications

The clinical outcomes that matter most in the NICU, and that a monitoring DT must ultimately predict, are characterized by two features that create structural statistical challenges: they are rare, and they are inconsistently defined. Among extremely preterm infants born before 28 weeks' gestation, necrotizing enterocolitis (NEC) occurs in approximately 9% [24]; late-onset sepsis in approximately 20% [24]; and severe intraventricular hemorrhage (IVH, grade III–IV) in approximately 14% [24]. Bronchopulmonary dysplasia (BPD) affects at least 40% of extremely preterm survivors depending on the definition applied; the 2018 NICHD workshop [25] proposed a severity-graded physiologic definition to address the limitations of earlier criteria, though BPD rates vary substantially across cohorts as a result of definitional heterogeneity. Mortality before NICU discharge among extremely preterm infants is approximately 22% overall in contemporary US academic cohorts [24], with variation that is steep across even single-week differences in gestational age at the threshold of viability, exceeding 50% at 23 weeks and falling below 10% by 28 weeks (Table 1). These outcomes are each a single, fixed binary endpoint per infant; a monitoring DT predicts them by refreshing its prediction as the physiologic predictor stream accumulates, a structure that is consistent with the repeated-measures predictor, subject-level binary outcome formulation that the validation methodology in Section 4 addresses.

These event rates have a direct consequence for DT model development that is frequently overlooked in the DT-IoT literature: at the single-center scale, the number of positive-outcome cases available to train a supervised classification model is often fewer than 20 per class, the regime in which misclassification error rates increase sharply and model stability degrades substantially, regardless of the learning algorithm employed (Figure 2) [11,26]. A hypothetical center admitting 60 extremely preterm infants per year at an approximately 9% NEC rate produces approximately five to six NEC cases annually, yielding a training dataset of perhaps 25–30 cases over five years of data collection. This is not a problem that can be solved through improved feature engineering or more complex model architectures; it is a fundamental constraint of single-center neonatal data that pushes NICU DT development inescapably toward multi-center architectures (introducing an additional level of data clustering) and that makes the methodological choices

in Sections 3 and 4 consequential rather than academic. Federated learning, discussed in Section 6, offers one architectural strategy for increasing effective sample size across this rare-event constraint without centralizing protected health information across institutions.

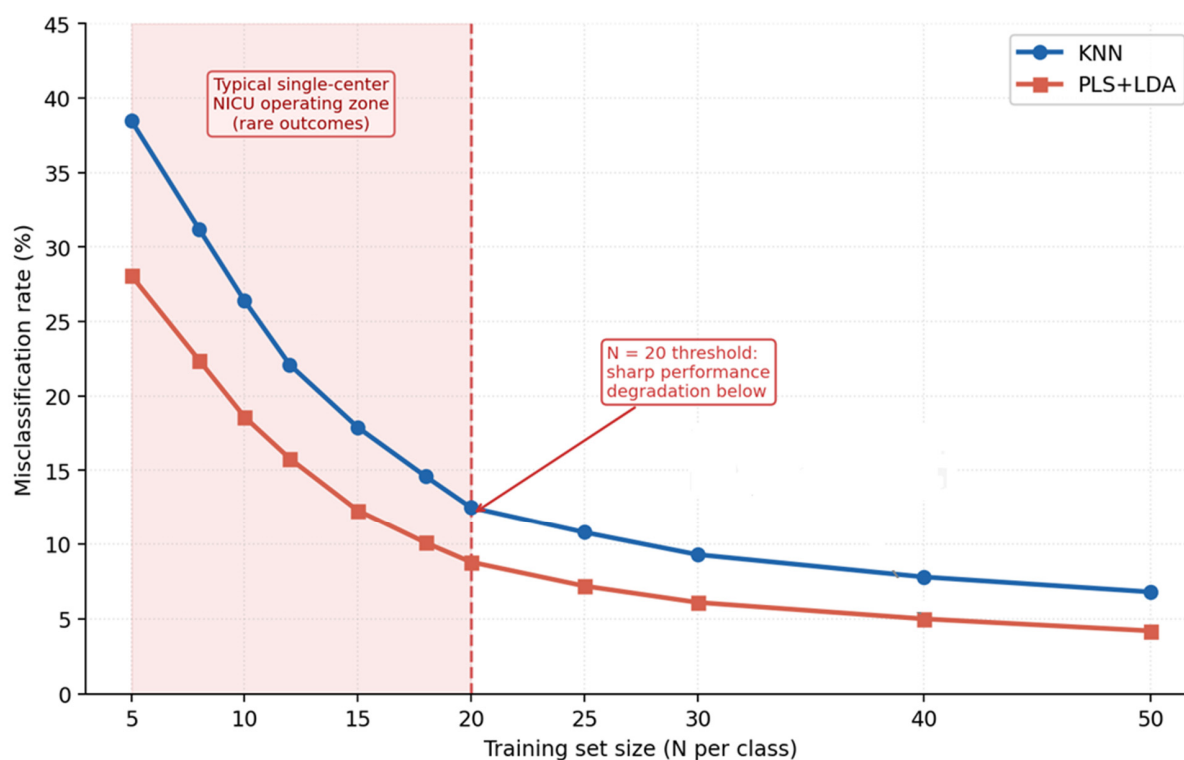


Figure 2. Misclassification rate as a function of training set size (N per class) for k-nearest neighbors (KNN) and partial least squares combined with linear discriminant analysis (PLS+LDA) classifiers [11,26]. The shaded zone represents the typical single-center NICU operating range for rare outcomes. Performance degrades sharply below N = 20 per class. The role of AI assistance in the preparation of this figure is disclosed in the Acknowledgments Section; the underlying data, axis labels, and interpretation were author-verified against the cited sources. Data are reproduced from a resampling simulation described in the author’s dissertation and in Hagan and Srivastav [26], using gene-expression microarray data (52 prostate cancer tumor samples and 50 normal prostate tissue samples). At each simulated sample size, a learning set of the specified size was drawn with replacement from a base set of 50 cases and 50 controls, defined as 100% sample size, with unselected observations serving as the test set; 100 such iterations were performed at each sample size, and each plotted value is the resulting mean misclassification rate for the best-performing gene-selection-method and gene-list-size combination at that sample size.

Table 1. Approximate outcome rates, prediction target type, data source, and primary validation challenge for principal NICU outcomes.

Outcome	Approx. Rate	Prediction Target	Data Source	Primary Validation Challenge
Necrotizing enterocolitis	~9%	Binary, subject-level	EHR, clinical diagnosis	Rare-event small effective N
Late-onset sepsis	~20%	Binary, subject-level	EHR, physiologic monitoring	Repeated-measures cluster leakage
Severe IVH (grade III–IV)	~14%	Binary, subject-level	Cranial ultrasound, EHR	Rare-event small effective N

Table 1. *Cont.*

Outcome	Approx. Rate	Prediction Target	Data Source	Primary Validation Challenge
Bronchopulmonary dysplasia	$\geq 40\%$ (definition-dependent)	Binary or severity-graded, subject-level	Respiratory support data, EHR	Outcome definition heterogeneity
Mortality before discharge	$\sim 22\%$ overall; steep gestational-age gradient	Binary, subject-level	EHR	Between-subject heterogeneity in baseline risk

2.3. What a Clinically Useful NICU DT Must Do

Following the taxonomy proposed by Drummond and Gonsard [7], neonatal DTs can be classified as either simulated or monitoring variants, a distinction that is not merely architectural but corresponds to a fundamental difference in statistical formulation. A monitoring DT maintains continuous synchronization with the physical twin through real-time IoT data feeds, updating its internal state as new physiologic information arrives, and is grounded in factual prediction models that estimate outcomes under observed conditions. A simulated DT is a one-time computational model of a patient state used for scenario testing or counterfactual prediction, estimating patient-specific outcomes under hypothetical interventions. The counterfactual formulation necessitates specialized identification assumptions (consistency, conditional exchangeability, positivity), estimation methods such as inverse probability weighting, and validation procedures beyond those of standard predictive modeling frameworks [27]. This review focuses on the monitoring DT, because its factual prediction task, mapping repeated-measures physiologic predictors to a clinical outcome, is the formulation for which the cluster-aware validation methodology described here is directly applicable. The simulated DT is acknowledged as clinically relevant but is treated as out of scope, since its counterfactual validation burden is a distinct problem.

For a monitoring DT, the core requirement is a prediction model whose estimated performance is trustworthy: discrimination and calibration that are accurately estimated, not optimistically inflated by methodological shortcuts. Because the predictors are repeated within-infant physiologic measurements, the principal threat to trustworthy performance estimation is the dependence structure of the data, which standard validation methods do not accommodate. A monitoring DT intended for bedside deployment must additionally satisfy operational requirements including calibration-driven alarm performance and prospective decision threshold validation (Section 4.4) and an HIPAA-compliant IoT data pipeline architecture (Section 6). The remainder of this review addresses these requirements in the order in which they arise in the DT development pipeline.

3. Supervised Learning Approaches for DT Construction

3.1. Framing the Prediction Problem

Constructing a monitoring DT from IoT data requires mapping a set of input features, derived from physiologic streams, EHR records, and demographic characteristics, to one or more clinical target outcomes. Depending on the application, that target may be a binary outcome (NEC or no NEC by day of life 30), a continuous physiologic trajectory (predicted arterial oxygen saturation over the next six hours), or a time-to-event endpoint (days to extubation readiness). Each formulation calls for a different family of supervised learning methods: classification models for binary and categorical outcomes, regression models for continuous targets, and survival or competing-risks models for time-to-event endpoints

with informative censoring. The choice among these formulations is not merely technical; it determines what a DT can and cannot represent.

In the NICU context, binary classification of adverse outcomes is the formulation most relevant to monitoring DT applications, and it is the formulation most severely challenged by the small effective sample sizes described in Section 2.2. The following subsections focus on this problem, though the methodological principles generalize to the other formulations.

3.2. Feature Selection Under High Dimensionality and Small Effective N

Feature selection, the process of identifying which inputs carry a predictive signal and which are noise, is a prerequisite for supervised learning on high-dimensional NICU IoT data, just as it is in genomic classification [9]. A continuous physiologic monitoring dataset for a single VLBW infant may contain hundreds of derived features: mean, variance, and distributional summaries of each waveform; cross-spectral coherence between channels; time-since-last-event features for discrete clinical data; and interaction terms between physiologic and clinical predictors. When the number of candidate features substantially exceeds the number of outcome events, a near-universal condition in single-center NICU data, standard maximum-likelihood estimation is undefined or severely overfit, and some form of regularization or pre-selection is essential.

Three broad strategies are available. Filter methods rank features by univariate association with the outcome (e.g., *t*-tests, Wilcoxon rank-sum statistics, or information-theoretic measures) and retain a top-*k* subset prior to model fitting. They are computationally efficient but ignore inter-feature dependencies and can be unstable: small perturbations of the training sample can produce substantially different feature rankings, a phenomenon well documented in high-dimensional biomedical settings [9,11,26,28]. Wrapper methods (e.g., recursive feature elimination, sequential forward selection) evaluate feature subsets by internal cross-validated model performance, making them sensitive to the specific learning algorithm and prohibitively expensive for large feature spaces. Embedded methods, most importantly LASSO (least absolute shrinkage and selection operator) and elastic net regularization, perform feature selection and model estimation simultaneously by penalizing the sum of absolute coefficient values, shrinking uninformative predictors exactly to zero [29]. For clinical prediction modeling, embedded methods such as LASSO are often favored when parsimony and clinical interpretability are priorities, as their tuning parameter can be selected by internal cross-validation and they produce sparse models well suited to deployment in clinical settings [17,30].

A critical but underappreciated property of all feature selection methods in small-sample settings is instability: the set of features selected varies substantially across bootstrap resamples of the training data when effective sample sizes are small [28]. Simulation evidence confirms that LASSO, while preferred among embedded methods, remains susceptible to this instability in the small-*N* regime characteristic of single-center NICU data [31]. This instability is not merely a theoretical concern. A neonatal DT whose predictive features vary substantially across resampled training iterations may be difficult to validate, update, or compare across institutions, as such variability can reflect sensitivity to sampling variation rather than a stable underlying signal. Stability-informed selection approaches (e.g., stability selection [32], which applies feature selection to many bootstrap subsamples and retains only features selected with high frequency) provide a principled framework for assessing the robustness of selected predictors and may improve reproducibility relative to single-pass selection. Given increasing recognition of the importance of model robustness and reproducibility in clinical prediction modeling, reporting measures of predictor stability is proposed here as a complement to existing reporting standards (e.g., calibration reporting in TRIPOD+AI, the AI-extended Transparent Reporting of a multivariable prediction model

for Individual Prognosis Or Diagnosis statement [19]) for NICU DT models. As a practical minimum, calibration-in-the-large and the calibration slope should be reported numerically alongside a flexible calibration curve constructed by locally weighted regression of the observed outcome on predicted probability; both are computable in standard statistical software (e.g., the rms and CalibrationCurves packages in R 4.6.1) without additional data collection.

3.3. Choice of Learning Algorithm for Sparse, High-Dimensional NICU Data

Given the feature selection landscape described above, the choice of supervised learning algorithm for NICU DT construction should be guided by three criteria: performance under small effective sample sizes, robustness to class imbalance, and interpretability sufficient to support clinical trust and regulatory review. Penalized logistic regression (e.g., LASSO or elastic net) satisfies all three reasonably well. Its performance at small N is competitive with more complex methods [10], its coefficient structure is transparent, and its output is a calibrated probability amenable to clinical decision thresholds. Evidence from a large-scale systematic review corroborates this preference: across clinical prediction modeling studies using tabular data, flexible machine learning methods demonstrate no average performance advantage over logistic regression [33]. Van Calster et al. [31] attribute this in part to the pervasive low signal-to-noise conditions that characterize clinical datasets, precisely the conditions that define single-center NICU data, where penalized regression is less susceptible to overfitting on irrelevant structure. Ensemble methods, particularly random forests and gradient boosting, offer greater modeling flexibility, but their advantage over penalized regression may diminish in smaller samples due to increased variance and risk of overfitting [10,11], and their black-box nature may create barriers to clinical adoption and regulatory clearance under FDA Digital Health frameworks [34]. Post hoc interpretability tools such as SHapley Additive exPlanations (SHAP) can quantify feature-level contributions to individual predictions and may partially mitigate these transparency barriers, providing model-agnostic explanations compatible with both ensemble and penalized regression methods.

Deep learning architectures, including recurrent neural networks (RNNs) and transformer-based models, are well suited to the temporal structure of NICU IoT streams in principle because they can model long-range dependencies within a physiologic time series that summary statistics discard. In practice, however, their performance depends on the availability of large training datasets [35], and they may be vulnerable to overfitting in smaller samples with sparse outcomes. Their appropriate role in NICU DT development is likely in transfer learning pipelines: pre-training on large unlabeled physiologic or EHR corpora to learn generalizable representations, followed by fine-tuning on labeled clinical outcomes, has shown improved performance, particularly in low-label settings [36,37], with additional regularization often required when adapting models to smaller NICU cohorts [38]. This architecture separates the data-volume requirement of representational learning (the unsupervised extraction of informative features from unlabeled input data) from the labeled-outcome requirement of clinical prediction, and is among the more promising near-term directions for the field. Domain-specific foundation models pretrained on pooled, multi-institutional critical care time series have begun to appear: self-supervised pretraining on physiologic and EHR streams pooled across institutions has produced early-stage critical care foundation models that outperform task-specific baselines when labeled outcome data are scarce [39,40]. These efforts remain rooted in adult critical care, are trained on substantially larger multi-site cohorts than any single-center NICU can supply, and have not yet been evaluated on neonatal physiologic streams or repeated-measures, subject-level binary outcomes of the kind addressed by this review; their relevance to NICU

DT development is accordingly a promising but empirically untested extension rather than an established practice. Federated pretraining across neonatal networks, discussed in Section 6 as a strategy for addressing single-center sample-size limitations, is a plausible route to assembling the multi-institutional data volume such foundation models require.

A related direction is multimodal AI, in which a single model jointly ingests physiologic time series alongside other data modalities, such as chest radiographs, cranial ultrasound images, or clinical narrative text, rather than treating each as a separate predictor stream. Multimodal architectures have shown gains over single-modality models in adult critical care prediction tasks by capturing cross-modal patterns that neither modality reveals alone. For NICU DTs, this raises the same identity-mediated leakage concern described for fixed and repeated-measures predictors in this review: an imaging or text modality tied to a specific infant is itself a quasi-identifier, and a validation strategy that partitions time series data by subject while inadvertently allowing the same infant's imaging or text data to appear in both training and validation folds would reintroduce exactly the leakage this review is intended to prevent. Multimodal NICU DT validation is accordingly noted as an open extension of the framework proposed here, not a solved problem.

3.4. Handling Temporal Dependence and Class Imbalance

Two structural features of NICU data require explicit methodological accommodation that generic machine learning pipelines do not provide by default. The first is temporal dependence: repeated observations within the same infant are serially correlated, violating the independence assumption underlying most standard learning algorithms. Ignoring this structure does not merely reduce efficiency, it invalidates standard cross-validation procedures (addressed in Section 4) and can produce spuriously confident predictions by treating autocorrelated observations as independent evidence. State-space models explicitly model within-infant temporal structure and are preferable to static snapshot approaches for monitoring DT construction [41]. Complementary approaches such as changepoint detection and mixed-effects extensions of penalized regression provide alternative frameworks for capturing temporal heterogeneity and clustering.

The second is class imbalance. For rare outcomes such as NEC (~9% incidence) or severe IVH (~14% incidence), a naive classifier that predicts the majority class for all observations achieves high accuracy but no clinical utility. Standard approaches to handling imbalance include cost-sensitive learning, which assigns higher misclassification penalties to the minority class during training, and resampling methods such as SMOTE (Synthetic Minority Over-Sampling Technique), which generates synthetic minority-class observations to rebalance the training set. A third approach, threshold adjustment, moves the classification decision boundary away from 0.5 toward a value that optimizes clinically relevant performance metrics such as sensitivity at a fixed specificity. Each approach involves tradeoffs, and the optimal strategy is outcome-specific and sample-size-dependent [42]. Importantly, any resampling of the training data to address imbalance must occur inside the cross-validation loop: applying SMOTE before splitting into folds leaks information from the validation set into the training set, producing optimistically biased performance estimates of exactly the type that makes clinical prediction models unreliable [19]. Van Calster et al. [31] document this and related cross-validation failures, including failure to freeze the fitted model before applying it to the test set and failure to repeat all preprocessing steps within each cross-validation fold, as among the most consequential sources of inflated apparent performance in published prediction models. Extensions of resampling to the longitudinal setting, which construct synthetic minority-class trajectories rather than synthetic single-timepoint observations, and cost-sensitive variants of tree-based ensembles tuned for time-series class imbalance, have been proposed as alternatives to

timepoint-level SMOTE; evaluating their performance specifically in the repeated-measures, rare-event NICU setting is noted as an open question in Section 8.2 rather than a settled recommendation here.

4. Cross-Validation and Model Evaluation in the NICU DT Context

4.1. Why Standard Cross-Validation Fails for NICU Data

Cross-validation is the principal tool for estimating how well a supervised learning model will perform on new, unseen data, and for selecting among competing models during development. In standard k-fold cross-validation, the dataset is partitioned into k non-overlapping folds; each fold serves in turn as the validation set while the remaining $k - 1$ folds form the training set, and performance estimates are averaged across folds. This procedure produces approximately unbiased performance estimates when observations are independent and identically distributed (i.i.d.), an assumption that is violated in fundamental ways by NICU IoT data.

Repeated physiologic measurements within the same infant are serially correlated: tomorrow's heart rate variability is predicted by today's, which was predicted by yesterday's. When temporally adjacent observations are split arbitrarily between training and validation folds, the validation set is not truly independent of the training data, as its observations remain statistically dependent on nearby training observations, so the estimated performance will be optimistically biased. This is the same phenomenon that inflates apparent performance in time-series forecasting when standard CV is naively applied [43], and it is directly relevant to monitoring DT development, where the input features are temporal summaries of physiologic streams.

Multicenter NICU data are clustered by infant and by center. Observations from the same infant share unmeasured individual characteristics (e.g., genetic background, antenatal exposures, institutional care practices) that make them more similar to each other than to observations from different infants. When a standard k-fold split places observations from the same infant in both the training and validation folds, the model is effectively evaluated on patients it has already seen in a different guise. In a multicenter dataset, the same problem scales up: a model trained and validated within a single institution has not been tested on the transportability that clinical deployment requires. Standard cross-validation is blind to both levels of clustering.

That this concern is more than theoretical is illustrated by recent NICU prediction practice. In a modeling competition using daily vital sign data from approximately 6000 NICU admissions [44], only one of five participating teams explicitly accounted for repeated-measures clustering during model evaluation, via leave-one-patient-out cross-validation. The model submitted by that team, a logistic regression with nonlinear terms and engineered heart rate variability features, achieved the highest area under the receiver operating characteristic curve on the held-out test data, outperforming submissions from the other four teams that used neural network, random forest, CatBoost, and XGBoost methods and applied standard k-fold cross-validation without patient-level partitioning. Comparable ambiguity regarding fold assignment is observed in published NICU prediction studies that apply k-fold CV to repeated daily observations without specifying whether folds are at the patient or observation level [45,46]. The methodological gap between cluster-aware best practice and routine analytic implementation in the neonatal prediction literature is therefore the rule rather than the exception, and is a direct risk for DT development efforts that draw on the same data structures and modeling traditions. The magnitude of optimism produced by observation-level cross-validation in this data structure has been quantified across a factorial parameter space in a recent simulation study by the present

author [17], on which the cluster-aware validation recommendations developed in this article are grounded.

Two distinct leakage axes therefore threaten validity in the monitoring DT setting, and they should be recognized as separate problems. The first is cluster leakage: observations from the same infant or center appearing in both training and validation partitions, addressed by cluster-aware partitioning as described in Section 4.2. The second is temporal look-ahead leakage: the use of predictor values recorded after the prediction time origin, which gives the model access to information that would not be available at the moment a prediction is made in deployment. A monitoring DT, which by definition generates predictions as physiologic data accumulate, is structurally susceptible to this second form of leakage, and a complementary form arises when modalities include outcome-correlated but non-causal signals, such as post-diagnostic laboratory values, post-treatment imaging reports, or clinical notes referencing the outcome of interest [47]. Both forms are mitigated by anchoring all features to a defined prediction time origin, by careful exclusion of post-diagnostic variables, and, where temporal autocorrelation is a concern, by the temporally ordered cross-validation designs described below. The cluster-aware and temporally ordered strategies are complementary: each addresses one leakage axis, and a monitoring DT validation pipeline must address both.

4.2. Appropriate Strategies for NICU DT Validation

Several cross-validation strategies offer progressively stronger protection against these sources of bias, at progressively higher computational and data-volume cost (Table 2). The strategies described here are not specific to neonatal data; they apply to any prediction model built on repeated-measures or clustered observations, and the NICU is used as the worked example. Standard naïve k-fold CV, as described above, is appropriate only for internal hyperparameter tuning on i.i.d. subsets, and should not be used to report final model performance for NICU DT models. Leave-one-out CV (LOOCV), the special limiting case of k-fold with k equal to the sample size, produces low-variance but often optimistically biased error estimates in small samples, and its discrete misclassification rate metric creates frequent ties that make it poorly suited to comparing competing classifiers [9]. Neither method adequately addresses the temporal or clustering structure of NICU data.

Table 2. Cross-validation strategy comparison for NICU DT models. Y = addresses; ~ = partially addresses; N = does not address.

CV Strategy	Temporal Autocorr.	Infant Clustering	Center Transport.	NICU DT Recommendation
Standard k-fold CV	N	N	N	Hyperparameter tuning only; do not use for final performance
Leave-one-out CV (LOOCV)	N	N	N	Optimistically biased in small samples; avoid for final reporting
Subject-level k-fold CV	~	Y	N	Minimum cluster-aware standard for repeated-measures predictors
Blocked time-series CV	Y	~	N	Preferred for single-center longitudinal models
Bootstrap time-series CV	Y	~	N	Preferred for rare outcomes (reduces variance)
Leave-one-center-out CV (LOCO-CV)	Y	Y	Y	Gold standard for multi-site deployment

Subject-level k-fold CV addresses cluster leakage by assigning all observations from any single infant to the same fold so that no infant contributes observations to both the training and validation partitions. This is the minimum cluster-aware standard for any prediction model developed from repeated within-subject measurements with a subject-level outcome. Leave-one-cluster-out CV, in which one cluster (an infant, or a center) is fully withheld at each iteration, is the limiting case of cluster-aware partitioning and provides the strongest protection against cluster leakage; leave-one-center-out CV (LOCO-CV) is its center-level form.

Blocked time-series CV addresses temporal autocorrelation by enforcing a temporal ordering constraint: training data always precede validation data in calendar time, and a gap may be inserted between the training and validation windows to prevent leakage of short-range autocorrelation. This approach correctly simulates prospective deployment: a model trained on NICU data from 2018–2021 is evaluated on 2022–2023 data, as it would be in practice. A refinement, bootstrap time-series CV, applies this blocked structure across many bootstrap resamples of temporally contiguous blocks rather than a single train–test partition, reducing the variance of the performance estimate without sacrificing temporal ordering. This approach is particularly valuable for rare outcomes, where a single blocked split may by chance assign most positive cases to either the training or validation window, producing unstable estimates. Both variants are preferred over standard CV for longitudinal NICU outcome models.

Leave-one-center-out CV is the most stringent strategy and the one most relevant to multi-site DT validation. Each center is held out in turn as the validation set, with the remaining centers comprising the training set, and performance is reported as the average across held-out centers. This procedure directly estimates the transportability of the model to a new institution, precisely the property that determines whether an NICU DT trained at an academic center with a 95th-percentile VLBW volume will perform adequately at a community hospital. LOCO-CV, or an equivalent multi-center hold-out design, is proposed as the minimum standard for reporting the final performance of any NICU DT model intended for multi-site deployment. Its computational cost is modest relative to bootstrap methods (it requires exactly C model fits for C centers) and its interpretability is high. While computational cost was once a practical barrier for iterative model fitting in large networks, penalized regression models of the type recommended in Section 3.3 fit in seconds on modern hardware, making LOCO-CV computationally accessible even for NICU research networks with $C = 20$ –30 participating centers.

A recent factorial simulation study of cross-validation strategies for prediction models with repeated-measures predictors and a subject-level binary outcome found that naive observation-level cross-validation can substantially overestimate model discrimination, with optimism in the area under the receiver operating characteristic (AUROC) curve ranging from +0.039 to +0.204 across 162 conditions. The magnitude of this bias increased monotonically with higher intraclass correlation, stronger temporal autocorrelation, fewer clusters, and lower event rates [18]. Subject-level k-fold cross-validation, in which all observations from any single infant are assigned to the same fold, produced AUROC estimates approximately concordant with leave-one-subject-out cross-validation across the conditions examined and is therefore recommended as the minimum cluster-aware standard for any NICU prediction model developed from repeated within-subject measurements [18]. To facilitate adoption of these strategies in neonatal prediction workflows, an open-source R package (cawCV) implementing the naive observation-level, subject-level, and leave-one-cluster-out cross-validation strategies described above is under development by the author [48]. A recent systematic review of 52 multimodal machine learning studies in predictive healthcare found that only 12% (6 studies) performed true external validation on

data from a different institution, time period, or acquisition setting [47], confirming that the gap between methodologically rigorous internal validation and demonstrated cross-site transportability is the dominant failure mode of contemporary clinical prediction modeling rather than an edge case.

Each strategy in Table 2 carries a corresponding limitation. Subject-level k-fold CV, the minimum standard, does not address temporal autocorrelation and can still produce optimistic estimates when within-subject correlation is strong relative to the number of subjects. Blocked and bootstrap time-series CVs address temporal structure but not cross-subject clustering, and are appropriate only when a single model is not also expected to generalize across centers. Leave-one-center-out CV provides the strongest protection on both axes but is the most computationally expensive of the four and requires data from multiple centers for implementation, which is unavailable during single-center model development. No single strategy in Table 2 is uniformly superior; the appropriate choice depends on which threats to validity are present in a given dataset and deployment setting.

A worked empirical example illustrates the magnitude of this optimism and demonstrates the application of NICU-DTVF Principle 2. In a cohort of 101 extremely low-birth-weight infants with daily oxygenation measurements and retinopathy of prematurity outcomes (Srivatsa et al., cited in [18]), naive observation-level 10-fold cross-validation produced an AUROC of 0.686 for severe ROP (15 events among 101 subjects), compared with 0.608 under leave-one-cluster-out cross-validation, an optimism of +0.078 AUROC units (95% percentile interval: +0.072, +0.082). For any ROP (48 events), the corresponding optimism was smaller but still material: +0.031 AUROC units (0.682 naive versus 0.650 LOCO). Subject-level 10-fold cross-validation closely approximated the LOCO reference in both cases (deviation ≤ 0.015), consistent with subject-level partitioning as the minimum adequate standard described above. The magnitude of optimism was larger for the rarer outcome (severe ROP, 15 events) than for the more common outcome (any ROP, 48 events), consistent with the general pattern that naive cross-validation optimism increases as the effective sample size for the outcome of interest decreases.

4.3. Performance Metrics Beyond Discrimination

AUROC has become the default performance metric for clinical prediction models, and it carries genuine meaning: it estimates the probability that a randomly selected positive subject receives a higher predicted probability than a randomly selected negative subject. For rare NICU outcomes, however, AUROC can be artificially high: a model that assigns modestly elevated probabilities to a small minority of true positives while correctly classifying the vast majority of true negatives can achieve an AUROC above 0.80 while providing limited clinical utility. The area under the precision–recall curve (AUPRC) is more informative for imbalanced outcomes and should be reported alongside AUROC for rare NICU endpoints [49].

Calibration, the agreement between predicted probabilities and observed event rates, is often the more clinically relevant metric for monitoring DT applications, and the one most frequently neglected [12]. A monitoring DT that is poorly calibrated will systematically over- or under-estimate event risk, degrading every downstream clinical decision that depends on its predicted probabilities. Calibration assessment should follow the framework recommended by Riley et al. [16] and Van Calster et al. [12]: calibration-in-the-large (mean predicted probability versus observed event rate), calibration slope (ideally 1.0; slopes less than 1.0 indicate overfitting), and a flexible calibration curve across the range of predicted probabilities reported graphically. Van Calster et al. [31] clarify that strong calibration, which requires correct risk estimates within every possible covariate pattern subgroup, is an unachievable standard in practice. The realistic and appropriate target is moderate

calibration, in which among all individuals assigned a predicted risk of X%, the observed event proportion approximates X%, evaluated via a flexible calibration plot. NICU DT models should be held to this achievable moderate calibration standard rather than an unachievable ideal.

Clinical utility assessment, currently absent from most NICU prediction model reports, should accompany discrimination and calibration metrics. Decision curve analysis (DCA), which evaluates net clinical benefit across a range of clinically plausible decision thresholds, provides an early approximation of potential clinical utility applicable to standard validation data, requiring no additional data collection [31]. DCA is appropriate for any NICU DT risk model for which a clinical decision threshold can be specified, and its inclusion in validation reports is recommended as a practical minimum for utility assessment. DCA provides an estimate of potential net benefit under assumed threshold preferences and does not substitute for prospective evaluation of the clinical intervention triggered by model-derived risk scores; definitive assessment of clinical impact requires a separate outcome study.

4.4. From Prediction to Action: Decision Thresholds and Clinical Deployment

A monitoring DT that outputs predicted probabilities must also specify how those probabilities translate into clinical action. This requires a decision threshold, a cutoff above which a prediction triggers an alert, initiates a monitoring escalation, or prompts clinical review. In standard binary classification, a default threshold of 0.5 is rarely clinically appropriate, particularly for rare adverse outcomes. For conditions such as NEC or severe IVH, the clinical cost of a missed case typically outweighs the burden of an unnecessary alert, justifying a lower threshold that improves sensitivity at an acceptable cost to specificity. The DCA framework introduced in Section 4.3 provides a principled basis for threshold selection: net benefit curves evaluated across clinically plausible decision thresholds make explicit the tradeoffs between false positives and false negatives and identify thresholds at which a model provides clinical utility relative to treating all or none [31]. Thresholds should be specified a priori and clinically justified, not derived post hoc from validation data alone.

The NICU environment poses a specific hazard for DT-derived risk alerts: alarm fatigue. Bedside monitoring systems in intensive care units are well documented to generate frequent alarms, the majority of which are clinically non-actionable, contributing to alarm fatigue among nursing staff. A poorly calibrated or over-triggered DT risk model compounds this problem, paradoxically reducing patient safety by degrading the signal-to-noise ratio of the clinical alert stream. Calibration quality directly determines alarm burden: a model that systematically overestimates risk will generate alerts far exceeding its true-positive rate, consuming nursing attention without clinical benefit. The moderate calibration target described in Section 4.3 is therefore not merely a statistical ideal but a patient safety requirement in the NICU monitoring context. Threshold selection should be evaluated prospectively against measured alarm rates, not only against held-out predictive performance, before any DT-derived alert system is deployed in clinical care.

The translation from prediction to automated action also determines a DT's regulatory classification under FDA Digital Health frameworks. A model that generates a risk probability for clinician review, without directly initiating treatment or triggering a protocol, occupies a lower regulatory risk tier than one whose outputs directly constrain clinical decisions or activate bedside device responses [34]. Regulatory oversight for AI/ML-based software increases with the level of clinical risk and intended use, with higher-risk applications requiring more rigorous premarket evaluation and ongoing post-market performance monitoring under the FDA's risk-based Software as a Medical Device

(SaMD) framework [34]. These regulatory considerations argue for a staged deployment strategy in which NICU DT risk models are first evaluated in passive-display mode, with alarm thresholds and clinical workflows developed and validated before any automated intervention capacity is introduced. The pipeline for responsible deployment, from model development through threshold validation to regulatory clearance, mirrors the framework described in Section 8.1 and should be treated as an integral component of DT development from the outset.

4.5. The Relation to Clinical Prediction Model Standards

The cross-validation and evaluation framework described here is aligned with the standards established for clinical prediction models by the TRIPOD+AI statement [19] and the Riley et al. sample size framework [16]. This alignment is intentional: a neonatal DT whose predictive component cannot meet TRIPOD+AI reporting standards is a DT whose methodology cannot be adequately scrutinized by peer reviewers, replicated by other investigators, or assessed for bias by regulatory bodies. The DT-IoT engineering literature has largely developed its own evaluation conventions, which emphasize reconstruction error and simulation fidelity metrics appropriate for physical systems but do not map cleanly onto clinical prediction validity. Adopting clinical prediction model standards as the methodological floor for NICU DT evaluation is not a constraint, it is the minimum condition for the field to produce evidence that clinical and regulatory audiences will find credible.

5. Synthetic Data and the Simulated DT: A Note on Scope

The methodology described in this review concerns the monitoring DT and its factual prediction task. A distinct line of work concerns the simulated DT and the generation of synthetic patient data, for example to augment underpowered randomized controlled trials with synthetic control arms, a possibility highlighted as a long-term motivation for neonatal DTs by Pammi et al. [6]. That application is acknowledged here as scientifically important but is outside the scope of the present review, because it raises a separate set of methodological problems: generative modeling of clustered, imbalanced physiologic data; counterfactual identification and estimation [27]; and the evaluation of whether synthetic data are fit for a given purpose. Synthetic data generated for trial augmentation must be evaluated for fidelity before use, and a tiered evaluation, progressing from marginal distributional fidelity, to joint and temporal dependency fidelity, and then to predictive fidelity assessed by the train-on-synthetic, test-on-real paradigm [50], is a reasonable structure for such evaluation. It is worth noting that the most stringent tier, predictive fidelity, is itself subject to the cluster-aware validation requirements described in Section 4: a train-on-synthetic, test-on-real evaluation conducted on repeated-measures neonatal data must partition the real hold-out at the infant level, because observation-level partitioning would inflate apparent predictive fidelity through exactly the optimism quantified in [18] and could allow a synthetic cohort to pass a predictive-fidelity check spuriously. The development of validated fidelity criteria for synthetic neonatal data, and of counterfactual validation methods for simulated DTs, is identified in Section 8 as a priority direction for future work but is not developed further here.

6. Privacy, Security, and IoT Architecture Considerations

The IoT data pipeline that feeds a neonatal DT, from the bedside device through the integration engine to the model training infrastructure, is governed at every stage by the Health Insurance Portability and Accountability Act (HIPAA) and its implementing regulations, which classify continuously recorded physiologic data linked to a patient

identifier as protected health information regardless of whether explicit identifiers are present [51]. This compliance obligation extends beyond the EHR to include waveform archives, ventilator data logs, and any derived feature dataset that retains sufficient granularity to permit re-identification. The re-identification risk is particularly acute in neonatal populations: combinations of quasi-identifiers can uniquely identify individuals in large populations [52], and this risk is amplified in high-dimensional clinical datasets where rare combinations of characteristics are more common [53]. In single-center NICU cohorts, variables such as gestational age, birthweight, institution, and admission timing may therefore render individuals effectively identifiable, necessitating suppression or generalization beyond standard de-identification approaches. Expert determination de-identification, which requires a statistical demonstration that re-identification risk falls below an acceptable threshold, is the appropriate standard for NICU DT development and should be documented and reported as part of any published methodology.

Federated learning offers an architectural solution to the tension between the multi-center data requirements of NICU DT development and the privacy constraints on centralizing patient data [54]. In a federated architecture, model parameters rather than raw patient data are transmitted between participating institutions and a central aggregator: each center trains on its own data locally, only model updates are shared, and the central model improves through aggregation across sites without any center's patient records leaving its institutional boundary. This approach maps naturally onto the structure of national neonatal research networks, where data governance frameworks for collaborative research already exist [55,56]. Federated learning is not without limitations in the NICU context, as statistical heterogeneity across centers can cause a federated model trained by simple gradient averaging to underperform a centrally trained model on any individual center's population, a problem known as client drift [57]. Edge computing, processing physiologic streams locally at the bedside to generate feature summaries before any data leave the device, provides a complementary layer of privacy protection for monitoring DT applications, though it introduces device security and software validation challenges that fall under FDA medical device regulations [34]. Not all IoT-derived data streams are equally feasible for real-time monitoring DT deployment. Continuously acquired waveform and device-derived signals (heart rate, SpO₂, respiratory rate, ventilator parameters) support real-time synchronization and are the natural input to a monitoring DT as defined in this review. Manually abstracted EHR fields, batch-processed laboratory results, and free-text clinical notes are presently feasible only for retrospective model development and periodic, rather than continuous, DT updating, because their acquisition and integration latency is incompatible with real-time synchronization. This distinction should inform which predictors a monitoring DT can incorporate at deployment versus which remain available only for retrospective model training and validation. Interoperability across the heterogeneous device and EHR systems that generate these streams remains a practical barrier to multi-institutional NICU DT deployment; standardized protocols such as Health Level Seven (HL7) and its Fast Healthcare Interoperability Resources (FHIR) specification are increasingly adopted for this purpose in healthcare DT applications generally [21], though device-level interoperability for high-frequency physiologic monitoring streams specifically remains less mature than EHR-level interoperability. Computational requirements for real-time inference at the bedside, sufficient processing capacity to generate updated predictions as physiologic data accumulate without introducing clinically meaningful latency, are a further practical constraint on monitoring DT deployment that is distinct from and additional to the statistical validation requirements that are the primary focus of this review.

7. Ethical and Regulatory Considerations

Neonatal DTs that support clinical decision-making fall within the scope of the FDA's Digital Health Center of Excellence and its guidance on artificial intelligence and machine learning-enabled medical devices [34]. A DT whose output directly informs treatment decisions would likely require premarket submission as Software as a Medical Device, with associated requirements for algorithmic transparency, performance reporting, and post-market surveillance [34]. Neonatal research conducted without the capacity for direct patient consent has always depended on parental proxy authorization, and DT development raises new questions about the scope of that proxy: parents who consent to their infant's clinical data being used for research may not have contemplated that consent as authorizing the construction of a computational model of their child. Pammi et al. [6] highlight the absence of formal consent guidance for pediatric digital health technologies as a key unresolved challenge, and this review concurs that the development of such frameworks is a genuine ethical prerequisite rather than a downstream implementation detail.

Equity concerns are equally pressing. DTs trained predominantly on data from large academic NICUs will encode the care patterns, population characteristics, and outcome distributions of those institutions, and may perform substantially worse when deployed at community hospitals serving different demographic populations. Prospective evaluation of DT performance across institutional types and patient demographic subgroups, with mandatory subgroup reporting analogous to that now expected in clinical trial publications, should be a condition of any regulatory submission or high-profile publication. This concern is not unique to neonatal applications; cross-specialty reviews of DT methodology in clinical research have similarly identified generalizability limitations as a primary equity risk when DTs are constructed from non-representative source populations, noting that rigorous validation techniques and data augmentation strategies are necessary to mitigate these constraints [58]. Economic evaluation of NICU DT deployment, whether the reduction in adverse outcomes and length of stay attributable to earlier risk detection offsets the infrastructure, integration, and surveillance costs of a validated monitoring DT, is presently absent from the neonatal DT literature and is identified as a research agenda item in Section 8.2. This evidence gap is not unique to neonatal applications; cost-effectiveness assessment is broadly underdeveloped across healthcare DT applications relative to the technical and clinical validation literature [21].

8. A Proposed Framework and Research Agenda

8.1. The NICU DT Validation Framework (NICU-DTVF)

The methodological considerations developed across Sections 3–7 converge on a structured framework for responsible NICU DT development, the NICU DT Validation Framework (NICU-DTVF). NICU-DTVF is best understood as a set of validation conditions that any trustworthy healthcare monitoring DT must satisfy, instantiated here for the neonatal repeated-measures setting. Cluster-aware cross-validation, implemented in the *cawCV* R package, is one component that addresses one of those conditions. As shown in Figure 1, key methodological requirements apply at distinct transitions in the DT pipeline. NICU-DTVF has four operative principles (Table 3). First, feature selection must be performed within the cross-validation loop and evaluated for stability across bootstrap resamples, with stability metrics reported alongside model performance. Second, the cross-validation strategy must match the data structure. Subject-level *k*-fold CV serves as the minimum cluster-aware standard, addressing optimism attributable to within-subject dependence [18]. Blocked or bootstrap time-series CV is recommended additionally when temporal autocorrelation constitutes a distinct source of leakage in longitudinal single-center models, and leave-one-center-out CV is the standard for any model

intended for multi-site deployment, with both the clustering and temporal leakage axes addressed. Third, performance reporting must include calibration assessment, calibration slope, calibration-in-the-large, and a flexible calibration curve, not discrimination alone. Fourth, decision thresholds and any automated alerting must be a priori specified and evaluated against threshold-level clinical performance, operating characteristics, and alarm burden before clinical deployment. The four principles differ in their current evidentiary and implementation status, a distinction stated here explicitly to avoid overinterpretation of the framework's maturity. The second principle, matching the cross-validation strategy to the clustering and temporal structure of the data, is both implemented in the *cawCV* R package and supported by simulation evidence quantifying the optimism produced when it is violated [18]; it is the most methodologically mature component of NICU-DTVF. The first and third principles, stability-assessed feature selection and calibration-inclusive performance reporting, are conceptual recommendations grounded in the broader clinical prediction model literature rather than components implemented in *cawCV* or empirically evaluated in the neonatal DT setting specifically. The fourth principle, a priori decision-threshold specification and alarm-burden evaluation, is a conceptual recommendation drawn from clinical deployment standards outside the prediction-modeling literature and has not been empirically evaluated in any NICU DT application to date. NICU-DTVF is accordingly best read as a set of validation conditions of varying evidentiary maturity, with cluster-aware cross-validation as its most developed component, rather than as a uniformly validated system. This framework has not been prospectively tested on a real multicenter NICU DT deployment, and such testing is identified as a priority item in the research agenda below (Section 8.2).

Table 3. NICU-DTVF operative principles, requirements, and evidentiary/implementation status.

Principle	Requirement	Evidentiary/Implementation Status
1. Feature selection stability	Perform selection within the CV loop; report stability metrics across bootstrap resamples	Conceptual recommendation, grounded in the broader clinical prediction model literature
2. CV strategy matched to data structure	Subject-level k-fold as minimum standard; blocked or bootstrap time-series CV for temporal autocorrelation; LOCO-CV for multi-site deployment	Implemented in <i>cawCV</i> ; supported by simulation and empirical evidence [18]
3. Calibration-inclusive performance reporting	Report calibration slope, calibration-in-the-large, and a flexible calibration curve, not discrimination alone	Conceptual recommendation, aligned with TRIPOD+AI [19]
4. A priori decision-threshold specification	Specify decision thresholds and alarm burden before deployment; evaluate prospectively	Conceptual recommendation; not yet empirically evaluated in a NICU DT application

This framework is deliberately aligned with TRIPOD+AI reporting standards [19] and the Riley et al. prediction model development guidelines [16], extending both to the DT context. Alignment with existing standards rather than invention of new ones is itself a methodological choice: the credibility of neonatal DT evidence will depend on whether it meets the evaluative standards that clinical and regulatory audiences already apply to prediction models. A DT that is evaluated by bespoke metrics intelligible only to the IoT engineering community will not influence clinical practice or regulatory policy. NICU-DTVF's original contribution, relative to TRIPOD+AI, is the specification of validation conditions for the repeated-measures, streaming-data structure of monitoring DT input, including cluster-aware cross-validation and streaming decision-threshold evaluation, requirements

TRIPOD+AI does not address because it presumes a single predictor observation per patient rather than a continuously updated physiologic stream. NICU-DTVF is similarly distinct from TRIPOD-Cluster [59], the existing reporting extension for prediction models developed or validated on clustered data: TRIPOD-Cluster addresses between-unit clustering, such as patients nested within hospitals or studies, whereas NICU-DTVF addresses within-subject clustering arising from repeated physiologic measurements over time, a structurally different dependence pattern that TRIPOD-Cluster's guidance does not cover and that motivates the cluster-aware cross-validation strategy central to this framework.

8.2. Prioritized Research Agenda

Against the framework above, the most consequential open methodological problems can be prioritized by three criteria: urgency (the degree to which the gap blocks current DT development), feasibility (the degree to which the problem is tractable without new data collection or additional infrastructure), and potential impact (the expected magnitude of improvement in DT validity or clinical trustworthiness if the problem were resolved). Figure 3 maps prioritized research agenda items across these dimensions, with urgency on the vertical axis, feasibility on the horizontal axis, and potential impact represented by bubble size. The nine problems shown represent the principal open methodological, infrastructural, and ethical challenges identified across Sections 3–7, grouped into three domains (machine learning methodology, infrastructure and regulatory, and ethics and equity). This set reflects the problems judged most consequential within the scope of the present review and is not intended to be exhaustive. Three machine learning methodology items emerge as high urgency and high feasibility, the appropriate starting point for the field. Establishing temporal and clustered CV standards for NICU outcome models requires no new data collection, only methodological comparison studies using existing multi-center cohorts such as those maintained by the NICHD Neonatal Research Network or the Canadian Neonatal Network. Developing stability assessment norms for feature selection in the NICU sample-size regime is similarly tractable using existing datasets and simulation methods. Adapting class-imbalance handling methods to NICU-specific outcome rates, with explicit evaluation of the requirement that resampling occur inside the cross-validation loop, can be accomplished within current single-center data environments.

Multicenter implementation of NICU-DTVF raises validation requirements beyond those addressed by the single-center cluster-aware and temporally ordered strategies described in Section 4.2. External validation (evaluating a model on data from institutions not used in model development), temporal validation (evaluating a model on data from a later calendar period than the development data), and geographic transportability (evaluating performance across health systems serving different patient populations) are each necessary conditions for a monitoring DT intended for deployment beyond the institution where it was developed. None is addressed by within-center cross-validation alone. Practical implementation at multicenter scale further requires a governance framework for cross-institutional data sharing (Section 6), harmonization of predictor definitions and measurement protocols across sites, and a two-level validation design that accounts for clustering of subjects within sites in addition to the within-subject repeated-measures clustering addressed here. Methodology for sample-size determination under this two-level clustering structure is an area in need of development as an extension to the single-level framework described in this review. These requirements are noted here as a priority research direction rather than resolved by the present review, which is scoped to single-center, within-subject validation.

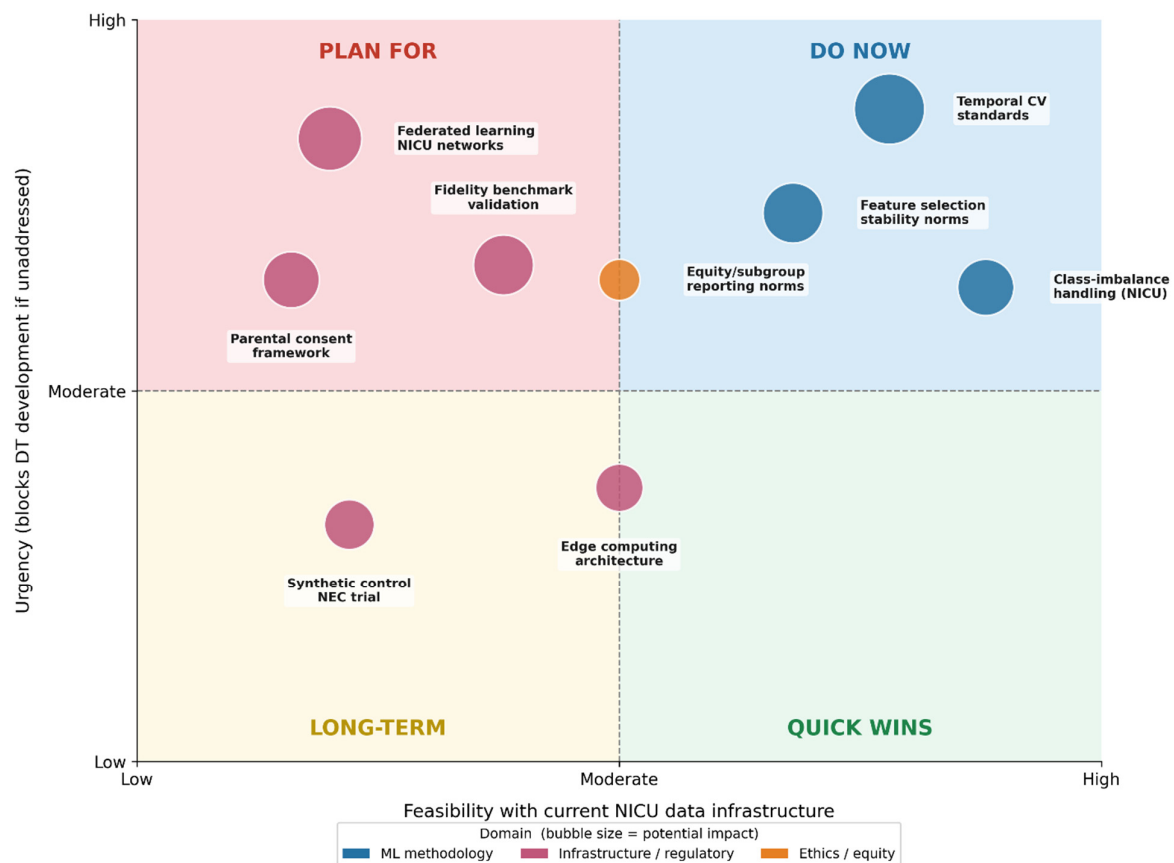


Figure 3. NICU DT research agenda: urgency × feasibility matrix. Nine prioritized open problems are mapped by urgency (*y*-axis), feasibility with current infrastructure (*x*-axis), and potential impact (bubble size). Quadrant labels indicate recommended prioritization strategy. Items in the upper-right quadrant (Do Now) represent the near-term critical path. Placements reflect a qualitative synthesis and are illustrative; see Section 8.2.

Items that are high urgency but lower feasibility require coordinated multi-center infrastructure that does not yet exist at scale, most notably the establishment of federated learning pipelines across NICU networks. Further items, including parental consent frameworks for DT use, equity and subgroup reporting norms, and the development of validated fidelity criteria and counterfactual validation methods for simulated DTs and synthetic data, occupy a moderate-urgency zone and will become more tractable as the higher-priority methodological problems are addressed. Two further items belong in this moderate-urgency zone and are noted explicitly here: privacy-preserving machine learning techniques, such as differential privacy and secure multi-party computation, as complements to the federated learning strategy discussed above for cross-institutional model development without centralizing protected health information, and synthetic data generation methods capable of producing realistic repeated-measures physiologic trajectories for simulated-DT and data-augmentation applications, which would need their own fidelity validation, as discussed in Section 5, rather than being assumed adequate by construction. It is recommended that the research community treat the cross-validation standards, feature selection stability, and class-imbalance handling problems as the near-term critical path, since progress on all three is prerequisite to credible NICU DT prediction models capable of supporting clinical deployment. The placement of each problem along these three dimensions reflects a qualitative synthesis of the analysis developed in this review rather than a formal scoring or consensus exercise. The exact position of any individual problem is accordingly illustrative, and a different investigator could reasonably

place individual items somewhat differently. The broad quadrant assignments, by contrast, follow from the methodological arguments developed in the preceding sections, and it is the quadrant-level prioritization, rather than the precise coordinates, that the figure is intended to convey.

8.3. A Coherent Methodological Program for Repeated-Measures Validation in Clinical Prediction

The validation strategies recommended in this review form part of a broader methodological program for clinical prediction modeling with repeated-measures predictors, in which several adjacent questions remain underdeveloped. Quantification of the optimism produced by naive cross-validation in this data structure has recently been addressed by a factorial simulation study by the present author [18], on which the cluster-aware validation recommendations in Section 4 are constructed. Implementation of these methods in open-source software (the cawCV R 4.6.1 package [48]) is in active development. Several adjacent methodological questions are not addressed in the present review and warrant further investigation. These include the behavior of calibration measures and the Brier score under within-subject dependence; sample-size determination for prediction models with repeated-measures predictors and a subject-level binary outcome; the effect of heterogeneity in within-subject correlation structure across predictors on validation performance; and the extension of cluster-aware validation to dynamic, time-updated prediction, in which a prediction is refreshed at successive landmark times as predictor data accumulate. The present review and the simulation study [18] on which it draws address the case of a single prediction per subject using the full observed predictor history; extension to dynamic, landmark-based prediction is one natural direction for future work. The cawCV package is intended to provide an extensible foundation for cluster-aware validation, to which methods addressing these questions may be added as they are developed.

These questions share a common structural feature: each concerns the translation of a standard prediction modeling tool, designed under the assumption of independent observations, to a setting in which observations are correlated within subjects. The cluster-aware cross-validation work [18] addresses this translation for the partitioning step of model evaluation. Calibration, sample size, dependence-structure heterogeneity, and dynamic prediction remain open analogues. Progress on these questions is a prerequisite for the kind of credentialed, transparently reported, and reproducibly validated NICU DT models that the framework described in Section 8.1 envisions, and for which TRIPOD+AI [19] and the Riley et al. sample-size framework [16] currently provide cross-sectional analogues without explicit guidance for the repeated-measures setting.

The convergence between this methodological program and the NICU DT validation framework described above is intentional. A neonatal DT cannot be validated more rigorously than the prediction models on which it is built, and prediction models that ignore the cluster structure of NICU IoT data will continue to overstate their generalization performance until cluster-aware validation methods become standard practice. The aim of this review is to make that convergence visible and to provide an explicit pathway by which methodological development in the validation of repeated-measures clinical prediction models can be coordinated with the clinical and translational ambitions of the neonatal DT vision. While that pathway is developed here in the neonatal setting, the validation methodology itself is general and transfers to any clinical prediction problem in which predictors are measured repeatedly within clustered units.

8.4. Limitations

This review has several limitations, consolidated here rather than left scattered across the preceding sections. First, NICU-DTVF itself is of mixed evidentiary maturity: as

stated in Section 8.1, its cluster-aware cross-validation component is implemented and empirically supported, while its feature selection, calibration-reporting, and decision-threshold components are conceptual recommendations not yet evaluated in an NICU DT application. Second, the framework has not been prospectively tested on a real multicenter NICU DT deployment; the multicenter validation requirements discussed in Section 8.2 remain a proposed research direction rather than a demonstrated solution. Third, the review is methodological rather than empirical in its primary contribution: with the exception of the worked example in Section 4.2, it does not report a new analysis of NICU data, and its recommendations are validated by simulation and by analogy to the broader clinical prediction model literature rather than by a dedicated neonatal DT deployment study. Fourth, the review's scope is deliberately restricted to the monitoring DT and to validation of the prediction model component; the simulated DT (Section 5), the engineering and regulatory infrastructure surrounding DT deployment, and multimodal data integration (Section 3.3) are each acknowledged but not resolved here. Finally, as in any methodological synthesis, the specific cross-validation, feature selection, and calibration recommendations reflect the state of the clinical prediction model and repeated-measures methodology literature as of this writing, and the open questions identified in Section 8.3 are expected to narrow, not close, as that literature develops.

9. Conclusions

The neonatal intensive care unit is a uniquely data-rich IoT environment whose clinical stakes are among the highest in the healthcare system. The outcomes determined there, from mortality to lifelong morbidity in the most vulnerable patients in medicine, are realized across an entire lifespan so that the potential health gains, measured for example in quality-adjusted life-years, are correspondingly large. The vision of monitoring digital twins that predict clinical events from continuously acquired physiologic data is compelling, scientifically grounded, and potentially transformative for a field constrained by the difficulty of acting on adverse outcomes before they become clinically manifest. What has been missing is not motivation but methodology.

This review has identified the specific methodological gaps that separate the monitoring DT vision from responsible clinical implementation: cross-validation approaches that violate the cluster and temporal structure of NICU IoT streams, feature selection instability under small effective sample sizes, and the relative neglect of calibration in model evaluation. For each gap, tractable solutions exist, drawn from the clinical prediction model literature and the time-series analysis literature, that have not yet been systematically applied in the neonatal DT context. The framework and research agenda proposed here are intended to provide a structured path from the current state to one in which neonatal DTs can be built with methodological rigor, evaluated against standards that clinical and regulatory audiences recognize, and deployed with equitable performance across the institutional diversity of neonatal care. The cluster-aware cross-validation methodology that anchors the validation component of NICU-DTVF [18], implemented in the open-source *cawCV* R package [48], together with the open methodological questions described in Section 8.3, comprises a coherent program of work on the validation of clinical prediction models with repeated-measures predictors, to which the present review contributes a conceptual synthesis and a framework for translation to the neonatal DT context. Realizing this program depends on progress along two further practical fronts beyond the statistical methodology emphasized here: the implementation challenges discussed in Sections 6 and 7, including interoperability across heterogeneous bedside and EHR systems, computational capacity for real-time bedside inference, and regulatory and consent frameworks suited to continuously updated risk prediction, and the multicenter research agenda in Section 8.2,

without which no NICU DT can be shown to generalize beyond the institution where it was developed. Methodological validity and practical deployability are jointly necessary; neither is sufficient on its own.

The limiting factor in realizing this vision is not computational capacity but methodological clarity and cross-disciplinary coordination between the neonatal clinical community, the biostatistics community, and the IoT and machine learning engineering community. This review is intended to contribute to that coordination.

Funding: This research received no external funding.

Institutional Review Board Statement: This review presents no new primary data. The clinical analyses referenced from prior work were conducted under Institutional Review Board approval as described in the original publications.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. The open-source R package cawCV [55], implementing the cluster-aware cross-validation methods described in this review, is permanently archived on Zenodo.

Acknowledgments: During the preparation of this manuscript, the author used Claude (Anthropic, San Francisco, CA, USA) to assist with reorganization of content, refinement of language, and drafting of figures. All methodological content, the selection of references, and the verification that cited references support the associated statements were directed and confirmed by the author. No AI tools were used for data analysis, statistical modeling, or the generation of results. The author has reviewed and edited all output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	artificial intelligence
AUC	area under the receiver operating characteristic curve
AUPRC	area under the precision-recall curve
BPD	bronchopulmonary dysplasia
CV	cross-validation
DCA	decision curve analysis
DT	digital twin
EHR	electronic health record
FDA	US Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level Seven
IoT	Internet of Things
IVH	intraventricular hemorrhage
KNN	k-nearest neighbors
LASSO	least absolute shrinkage and selection operator
LOCO-CV	leave-one-center-out cross-validation
LOOCV	leave-one-out cross-validation
ML	machine learning
NEC	necrotizing enterocolitis
NICHD	Eunice Kennedy Shriver National Institute of Child Health and Human Development
NICU	neonatal intensive care unit

PHI	protected health information
PLS+LDA	partial least squares with linear discriminant analysis
RNN	recurrent neural network
ROP	retinopathy of prematurity
SaMD	Software as a Medical Device
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-Sampling Technique
VLBW	very low birthweight

References

- Grievens, M.; Vickers, J. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In *Transdisciplinary Perspectives on Complex Systems*; Kahlen, F.-J., Flumerfelt, S., Alves, A., Eds.; Springer: Cham, Switzerland, 2017; pp. 85–113.
- Tao, F.; Zhang, H.; Liu, A.; Nee, A.Y.C. Digital twin in industry: State-of-the-art. *IEEE Trans. Ind. Inform.* **2019**, *15*, 2405–2415. [\[CrossRef\]](#)
- Fuller, A.; Fan, Z.; Day, C.; Barlow, C. Digital twin: Enabling technologies, challenges and open research. *IEEE Access* **2020**, *8*, 108952–108971. [\[CrossRef\]](#)
- Björnsson, B.; Borrebaeck, C.; Elander, N.; Gasslander, T.; Gawel, D.R.; Gustafsson, M.; Jörnsten, R.; Lee, E.J.; Li, X.; Lilja, S.; et al. Digital twins to personalize medicine. *Genome Med.* **2020**, *12*, 4. [\[CrossRef\]](#) [\[PubMed\]](#)
- Corral-Acero, J.; Margara, F.; Marciniak, M.; Rodero, C.; Loncaric, F.; Feng, Y.; Gilbert, A.; Fernandes, J.F.; Bukhari, H.A.; Wajdan, A.; et al. The ‘DT’ to enable the vision of precision cardiology. *Eur. Heart J.* **2020**, *41*, 4556–4564. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pammi, M.; Shah, P.S.; Yang, L.; Hagan, J.; Aghaeepour, N.; Neu, J. Digital twins and synthetic patient data: Can they empower clinical trials in children? *Lancet Digit. Health* **2025**, *7*, 100851. [\[CrossRef\]](#) [\[PubMed\]](#)
- Drummond, C.; Gonsard, M. Definitions and characteristics of patient digital twins being developed for clinical use: Scoping review. *J. Med. Internet Res.* **2024**, *26*, e58504. [\[CrossRef\]](#) [\[PubMed\]](#)
- Moorman, J.R.; Carlo, W.A.; Kattwinkel, J.; Schelonka, R.L.; Porcelli, P.J.; Navarrete, C.T.; Bancalari, E.; Aschner, J.L.; Walker, M.W.; Perez, J.A.; et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: A randomized trial. *J. Pediatr.* **2011**, *159*, 900–906.e1. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ambroise, C.; McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6562–6566. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
- Hagan, J.L. Comparison of Supervised Learning Methods for Classification of Microarray Data. Doctoral Dissertation, Department of Biostatistics, Tulane University, New Orleans, LA, USA, 2012.
- Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W.; on behalf of the Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lee, Y.-H.; Nam, T.; Cho, D.-S.; Kim, W.-T. LLM-Based Adaptive Control Code Generation Framework with Digital Twin-Integrated Verification for Heterogeneous Robot Systems. *Appl. Sci.* **2026**, *16*, 3883. [\[CrossRef\]](#)
- di Benedetto, M.; Randazzo, V.; Lidozzi, A.; Accetta, A.; Ghione, G.; Solero, L.; Cirrincione, G.; Pasero, E.G.A. Enhanced Neural Real-Time Digital Twin for Electrical Drives. *Appl. Sci.* **2026**, *16*, 3955. [\[CrossRef\]](#)
- National Academies of Sciences, Engineering, and Medicine. *Foundational Research Gaps and Future Directions for Digital Twins*; National Academies Press: Washington, DC, USA, 2024. [\[CrossRef\]](#) [\[PubMed\]](#)
- Riley, R.D.; Ensor, J.; Snell, K.I.E.; Harrell, F.E.; Martin, G.P.; Reitsma, J.B.; Moons, K.G.M.; Collins, G.; van Smeden, M. Calculating the sample size required for developing a clinical prediction model. *BMJ* **2020**, *368*, m441. [\[CrossRef\]](#) [\[PubMed\]](#)
- Steyerberg, E.W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed.; Springer: New York, NY, USA, 2019.
- Hagan, J.L. Quantifying the Optimism of Naive Cross-Validation for Binary Outcome Prediction with Repeated-Measures Predictors: A Simulation Study and Clinical Illustration. *medRxiv* **2026**. [\[CrossRef\]](#)
- Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [\[CrossRef\]](#) [\[PubMed\]](#)

20. Laubenbacher, R.; Sluka, J.P.; Glazier, J.A. Using digital twins in viral infection. *Science* **2021**, *371*, 1105–1106. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Khoshfekar Rudisari, H.; Tseng, B.; Zhu, H.; Song, L.; Gu, C.; Roy, A.; Irajizad, E.; Butner, J.; Long, J.; Do, K.-A. Digital twins in healthcare: A comprehensive review and future directions. *Front. Digit. Health* **2025**, *7*, 1633539. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Soul, J.S.; E Hammer, P.; Tsuji, M.; Saul, J.P.; Bassan, H.; Limperopoulos, C.; Disalvo, D.N.; Moore, M.; Akins, P.; Ringer, S.; et al. Fluctuating pressure-passivity is common in the cerebral circulation of sick premature infants. *Pediatr. Res.* **2007**, *61*, 467–473. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Sward-Comunelli, S.L.; Mabry, S.M.; Thibeault, D.W.; Truog, W.E. Ventilator support after surfactant therapy for respiratory distress syndrome: Patterns of use. *J. Perinatol.* **1997**, *17*, 296–302.
24. Bell, E.F.; Hintz, S.R.; Hansen, N.I.; Bann, C.M.; Wyckoff, M.H.; DeMauro, S.B.; Walsh, M.C.; Vohr, B.R.; Stoll, B.J.; Carlo, W.A.; et al. Mortality, in-hospital morbidity, care practices, and 2-year outcomes for extremely preterm infants in the US, 2013–2018. *JAMA* **2022**, *327*, 248–263. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Higgins, R.D.; Jobe, A.H.; Koso-Thomas, M.; Bancalari, E.; Viscardi, R.M.; Hartert, T.V.; Ryan, R.M.; Kallapur, S.G.; Steinhorn, R.H.; Konduri, G.G.; et al. Bronchopulmonary dysplasia: Executive summary of a workshop. *J. Pediatr.* **2018**, *197*, 300–308. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Hagan, J.L.; Srivastav, S.K. Performance of Partial Least Squares + Linear Discriminant Analysis versus k-Nearest Neighbors for Validation Set Classification of Cancer DNA Microarray Data. *Biostat. Biom. Open Access J.* **2019**, *9*, 555752. [\[CrossRef\]](#)
27. Boyer, C.B.; Dahabreh, I.J.; Steingrimsson, J.A. Estimating and evaluating counterfactual prediction models. *Stat. Med.* **2025**, *44*, e70287. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Meinshausen, N.; Bühlmann, P. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2010**, *72*, 417–473. [\[CrossRef\]](#)
29. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [\[CrossRef\]](#)
30. Pavlou, M.; Ambler, G.; Seaman, S.; De Iorio, M.; Omar, R.Z. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat. Med.* **2016**, *35*, 1159–1177. [\[PubMed\]](#)
31. Van Calster, B.; van Smeden, M.; van Amsterdam, W.; Coemans, M.; Wynants, L.; Steyerberg, E.W. The enemies of reliable and useful clinical prediction models. *Annu. Rev. Stat. Appl.* **2026**, *13*, 465–492. [\[CrossRef\]](#)
32. Shah, R.D.; Samworth, R.J. Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2013**, *75*, 55–80. [\[CrossRef\]](#)
33. Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; Van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **2019**, *110*, 12–22. [\[CrossRef\]](#) [\[PubMed\]](#)
34. US Food and Drug Administration. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*; FDA: Silver Spring, MD, USA, 2022.
35. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in health and medicine. *Nat. Med.* **2022**, *28*, 31–38. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Pellegrini, C.; Navab, N.; Kazi, A. Unsupervised pre-training of graph transformers on patient population graphs. *Med. Image Anal.* **2023**, *89*, 102895. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Mataraso, S.J.; Espinosa, C.A.; Seong, D.; Reincke, S.M.; Berson, E.; Reiss, J.D.; Kim, Y.; Ghanem, M.; Shu, C.-H.; James, T.; et al. A machine learning approach to leveraging electronic health records for enhanced omics analysis. *Nat. Mach. Intell.* **2025**, *7*, 293–306. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Harutyunyan, H.; Khachatryan, H.; Kale, D.C.; Steeg, G.V.; Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **2019**, *6*, 96. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Jagd, K.N.; DeVries, R.; Winther, O. Towards Self-Supervised Foundation Models for Critical Care Time Series. *arXiv* **2025**, arXiv:2509.19885. [\[CrossRef\]](#)
40. Burger, M.; Chopard, D.; Londschie, M.; Sergeev, F.; Yeche, H.; Kuznetsova, R.; Faltys, M.; Gerdes, E.; Leshetkina, P.; Buhlmann, P.; et al. A Foundation Model for Intensive Care: Unlocking Generalization across Tasks and Domains at Scale. *medRxiv* **2025**. [\[CrossRef\]](#)
41. Durbin, J.; Koopman, S.J. *Time Series Analysis by State Space Methods*, 2nd ed.; Oxford University Press: Oxford, UK, 2012.
42. Guo, H.; Li, Y.; Shang, J.; Gu, M.; Huang, Y.; Gong, B. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [\[CrossRef\]](#)
43. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [\[CrossRef\]](#)
44. Sullivan, B.A.; Moreira, A.G.; McAdams, R.M.; Knake, L.A.; Husain, A.; Qiu, J.; Mudireddy, A.; Majeedi, A.; Shalish, W.; Lake, D.E.; et al. Comparing machine learning techniques for neonatal mortality prediction: Insights from a modeling competition. *Pediatr. Res.* **2025**, *98*, 405–411. [\[PubMed\]](#)
45. Niestroy, J.C.; Moorman, J.R.; Levinson, M.A.; Manir, S.A.; Clark, T.W.; Fairchild, K.D.; Lake, D.E. Discovery of signatures of fatal neonatal illness in vital signs using highly comparative time-series analysis. *npj Digit. Med.* **2022**, *5*, 6. [\[CrossRef\]](#) [\[PubMed\]](#)

46. Song, W.; Jung, S.Y.; Baek, H.; Choi, C.W.; Jung, Y.H.; 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**, *8*, e15965. [[CrossRef](#)] [[PubMed](#)]
47. Vavekanand, R.; Kumar, T.; Kumar, S.; Kumar, G.; Laghari, A.A. Multimodal Machine Learning Approaches in Predictive Healthcare Analytics: A Comprehensive Survey. *Arch. Comput. Methods Eng.* **2026**, *33*, 7667–7691. [[CrossRef](#)]
48. Hagan, J.L. cawCV: Cluster-Aware Cross-Validation for Repeated-Measures Predictors with Binary Outcomes [Software]. *Zenodo* **2026**. [[CrossRef](#)]
49. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [[CrossRef](#)] [[PubMed](#)]
50. Esteban, C.; Hyland, S.L.; Rätsch, G. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv* **2017**, arXiv:1706.02633.
51. US Department of Health and Human Services. HIPAA Security Rule. 45 CFR Parts 160 and 164. *Fed. Regist.* **2003**, *68*, 8334–8381.
52. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2002**, *10*, 557–570. [[CrossRef](#)]
53. El Emam, K.; Jonker, E.; Arbuckle, L.; Malin, B. A systematic review of re-identification attacks on health data. *PLoS ONE* **2011**, *6*, e28071. [[CrossRef](#)] [[PubMed](#)]
54. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Agüera y Arcas, B. Communication-efficient learning of deep networks from decentralized data. *Proc. Mach. Learn. Res.* **2017**, *54*, 1273–1282.
55. Horbar, J.D.; Soll, R.F.; Edwards, W.H. The Vermont Oxford Network: A community of practice. *Clin. Perinatol.* **2010**, *37*, 29–47. [[CrossRef](#)] [[PubMed](#)]
56. Bell, E.F.; Stoll, B.J.; Hansen, N.I.; Wyckoff, M.H.; Walsh, M.C.; Sánchez, P.J.; Rysavy, M.A.; Gabrio, J.H.; Archer, S.W.; Das, A.; et al. Contributions of the NICHD Neonatal Research Network’s Generic Database to documenting and advancing the outcomes of extremely preterm infants. *Semin Perinatol.* **2022**, *46*, 151635. [[CrossRef](#)] [[PubMed](#)]
57. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Smola, A.; Smith, V. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Res.* **2020**, *2*, 429–450.
58. Akbarialiabad, H.; Pasdar, A.; Murrell, D.F.; Mostafavi, M.; Shakil, F.; Safaei, E.; Leachman, S.A.; Haghighi, A.; Tarbox, M.; Bunick, C.G.; et al. Enhancing randomized clinical trials with digital twins. *npj Syst. Biol. Appl.* **2025**, *11*, 110. [[CrossRef](#)] [[PubMed](#)]
59. Debray, T.P.A.; Collins, G.S.; Riley, R.D.; Snell, K.I.E.; Van Calster, B.; Reitsma, J.B.; Moons, K.G.M. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. *BMJ* **2023**, *380*, e071018. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.