Critical Care

https://doi.org/10.1186/s13054-025-05765-1

Article in Press

Evaluating the effect of heart and respiratory rate measurement errors on the ability to predict the outcome of high flow nasal cannula therapy: a multi-centre study

Received: 26 August 2025

Accepted: 12 November 2025

Published online: 22 November 2025

Cite this article as: Yu H., Saffaran S., Tonelli R. *et al.* Evaluating the effect of heart and respiratory rate measurement errors on the ability to predict the outcome of high flow nasal cannula therapy: a multi-centre study. *Crit Care* (2025). https://doi.org/10.1186/s13054-025-05765-1

Hang Yu, Sina Saffaran, Roberto Tonelli, John G. Laffey, Qingchen Zhang, Antonio M. Esquinas, Lucas Martins Lima, Letícia Kawano-Dourado, Israel S. Maia, Alexandre Biasi Cavalcanti, Enrico Clini & Declan G. Bates

Evaluating the effect of heart and respiratory rate measurement errors on the ability to predict the outcome of high flow nasal cannula therapy: a multi-centre study

Hang Yu[1], Sina Saffaran[1], Roberto Tonelli[2,3], John G. Laffey[4,5], Qingchen Zhang[6], Antonio M. Esquinas[7,8], Lucas Martins de Lima[9], Letícia Kawano-Dourado[9], Israel S. Maia[9], Alexandre Biasi Cavalcanti[9], Enrico Clini[2,3,*], and Declan G. Bates[1]

[1] School of Engineering, University of Warwick, Coventry CV4 7AL, UK.
[2] Department of Medical and Surgical Sciences of Adult and Mother-Child SMECHIMAI, University of Modena Reggio-Emilia, Modena, Italy.
[3] University Hospital of Modena Policlinico, Respiratory Diseases Unit, Modena, Italy
[4] Anaesthesia and Intensive Care Medicine, Galway University Hospitals, Galway, Ireland.
[5] School of Medicine, University of Galway, Galway, Ireland.
[6] School of Computer Science and Technology, Hainan University Haikou, China.
[7] Intensive Care Unit, Hospital Morales Meseguer, Murcia, Spain.
[8] NIV-ICM Research Group, Biomedical Research Institute of Murcia (IMIB-Pascual Parrilla), Murcia, Spain.
[9] Hcor Research Institute, Hcor Hospital, Rua Desembargador Eliseu Guilherme, 200 Paraíso, São Paulo 04004-030, Brazil.

***Corresponding author**: Enrico Clini: enrico.clini@unimore.it

Abstract

Background: The respiratory rate oxygenation (ROX) index, and machine learning (ML) models, are promising approaches to help clinicians identify earlier those patients at risk of failing high flow nasal cannula (HFNC) therapy. Respiratory rate (RR) and heart rate (HR) are key inputs to these models, but their measurement in a hospital environment may be subject to significant errors. The effect of these errors on the accuracy of HFNC outcome predictions is currently unknown.

Methods: We evaluated the capability of a recently-proposed ML model called Tabular Prior-data Fitted Network (TabPFN), a range of standard ML models, and the ROX index and its variants, to predict the outcome of HFNC therapy using measurements made within the first 2 hours of treatment in patients with acute hypoxemic respiratory failure. 596 AHRF patients receiving HFNC (456 successes, 140 failures) from the RENOVATE trial in Brazil were used for model training. External validation was performed on a dataset on 241 AHRF patients (156 successes, 85 failures) from Italy and the US. During training and testing, we replicated RR and HR measurement errors that were consistent with those recorded in previously published studies employing 30-second and 15-second manual counting time-windows, respectively, and employed bootstrapping and Monte Carlo simulation to evaluate their effects on the accuracy of outcome predictions.

Results: The TabPFN model was more affected by the RR and HR measurement errors, but still provided more accurate predictions of HFNC outcome (Mean [95% CI] Accuracy 0.79 [0.73-0.84], AUC 0.86 [0.82-0.89]

in external validation) than the ROX index and its variants (Accuracy 0.71 [0.68-0.75], AUC 0.78 [0.75-0.80]). Augmenting patient datasets with arterial blood gas measurements further improved the performance and robustness of the TabPFN model, but not the ROX index.

Conclusions: In this multi-centre study, the recently introduced TabPFN ML model outperformed currently available methods for predicting the outcome of HFNC therapy even when realistic levels of measurement errors were included in the clinical data on RR and HR.

The predictive performance of this ML model can be further improved by minimizing measurement errors using more advanced monitoring, and/or by additionally using arterial blood gas measurements.

Introduction

Respiratory rate (RR) and heart rate (HR) are vital signs widely utilized in clinical assessments, and are integral components of numerous risk prediction scoring systems [1]. In patients with acute hypoxemic respiratory failure (AHRF) receiving high-flow nasal cannula (HFNC), early identification of individuals at high risk of failure can enable timely escalation of care and potentially improve outcomes [2-4]. Clinical indices that use RR and HR, such as the respiratory rate oxygenation (ROX) index and its variants have been proposed as potential predictors of HFNC failure [5-8]. More recently, data-driven approaches utilizing machine learning (ML) models have also shown potential to enhance the reliability of outcome prediction for HFNC therapy [9].

However, a critical but often overlooked factor in assessing the reliability of both clinical indices and ML models is the quality of their input data, particularly RR and HR, where measurement errors are common but rarely accounted for [10,11,12]. In clinical practice, RR is typically measured in patients receiving HFNC by manually counting chest-wall movements over a 30 second sample duration and extrapolating to breaths per minute, a time-consuming and error-prone process [13]. Prior studies have shown that such short-duration RR measurements are prone to error and variability, with a mean interquartile error range of 3.4 breaths per minute in a cohort of 25 patients reported by Drummond et al. [14].

Similarly, although HR will typically be measured in the ICU via electrocardiography, pulse oximetry or other automatic monitoring methods, in general wards or emergency rooms where HFNC is often used

manual counting by radial palpation is still common. Although most nursing textbooks recommend a 60 second time-window for accurate pulse-counting [15], one study suggested that more than 75% of nurses used shorter (15 or 30 second) windows [16], increasing the potential for inaccuracies due to counting errors and/or periodicity and instability of the heartbeat sequence. Novel wearable optical heart rate sensors have also been reported to produce significant errors [17].

Given that RR and HR are key components of both the standard ROX indices and recently proposed ML models, it is important to understand how underlying measurement errors might impact their performance. However, to our knowledge the sensitivity of HFNC outcome predictors to inaccuracies in these measurements has not so far been systematically investigated.

Here, we rigorously quantify the effects of RR and HR measurement errors, at levels likely to be encountered in current clinical practice, on the predictive accuracy of the ROX index (and its variants), of a recently introduced ML model called Tabular Prior-data Fitted Network (TabPFN) [18], and of a range of standard ML models. The new TabPFN model, which uses in-context learning [19], the mechanism underlying the unprecedented performance of large language models, has been specifically developed for the kind of small-to-medium-sized datasets that are currently available on patients receiving HFNC therapy, and has been shown in a recent study to significantly out-perform other methods for predicting the outcome of non-invasive ventilation [20].

Methods

Study cohorts

In this retrospective multi-centre analysis, clinical data were obtained from publicly accessible databases and previously published research studies (Additional File: Figure S3). The internal training cohort used for model training and cross-validation included 596 patients (456 HFNC successes, 140 failures) diagnosed with AHRF, derived from the recent RENOVATE trial, involving 33 hospitals across Brazil [21]. The external validation cohort comprised a total of 241 patients (156 HFNC successes and 85 failures). This cohort consisted of patient data from two pilot studies conducted at the Respiratory Intensive Care Unit of the University Hospital of Modena, Italy (184 patients, 116 successes, 68 failures) [22, 23], the publicly available MIMIC-IV database from the US (38 patients, 21 successes and 17 failures) [24], and the eICU database from the US (19 patients, all successes) [25] (see Additional File: Figure S4 for data extraction procedure). All patients in these studies met the following admission criteria: adult patients (age >18) with *de novo* AHRF who had failed standard oxygen therapy and were evaluated for escalation to HFNC therapy (Additional File: Table S1) - for a comprehensive overview of the study data characteristics across internal training and external validation cohorts, see Table 1.

Data collection for the Modena cohorts was conducted under the approval of the Area Vasta Emilia Nord Ethics Committee (protocol 266/2016/OSS/AOUMO), and their use for further analysis is covered by the CORALINE protocol (165/2024/SPER/AOUMO, SIRER ID 7354). The

RENOVATE data was analyzed under the ethical approvals granted for the original studies (Hcor Ethics Committee No. 2.888.697, approval date [12/09/2018]). Per the institutional policies and regulatory guidance, secondary analyses of the original, de-identified datasets do not require a new ethics submission when conducted with participation by the original investigators under the scope of the initial approval.

Definition and Measurement of Clinical Indices

The clinical indices for predicting the outcome of HFNC therapy evaluated were the following: ROX = $\frac{SpO_2}{FiO_2 \cdot RR}$, mROX = $\frac{P_aO_2}{FiO_2 \cdot RR}$, ROX-HR = $\frac{100 \cdot SpO_2}{FiO_2 \cdot RR \cdot HR}$, and mROX-HR = $\frac{100 \cdot P_aO_2}{FiO_2 \cdot RR \cdot HR}$.

HFNC failure across all cohorts was defined as the need for non-invasive or invasive mechanical ventilation (MV), or death, within 24 hours of HFNC initiation (see Figure 1B). The patient measurements were taken at two timepoints: baseline (T0), defined as within 6 hours prior to or at the initiation of HFNC, and a follow-up timepoint (T1), defined as 1-2 hours after HFNC initiation. No patients had more than one measurement available in each time window. Optimal thresholds for each index were determined using univariate logistic regression models fitted on the development cohort, with the threshold determined based on maximizing balanced accuracy. Missing values (summarised in Additional File: Figure S2) were first addressed using forward-fill imputation, and any remaining were subsequently handled through k-nearest neighbours imputation (Figure 1B).

Machine learning models

We developed and evaluated a novel machine learning model to predict HFNC outcome using the recently introduced TabPFN algorithm [18], a pre-trained Transformer specifically designed for classification tasks on small tabular datasets without the need for hyperparameter tuning. In-context learning was employed to encode the provided training data directly within its attention mechanism to generate predictions. This approach allows the model to leverage its extensive prior knowledge, obtained during pre-training on diverse synthetic datasets, in combination with context-specific training examples to make accurate predictions for new data points (Additional File: TabPFN details). The process for data preprocessing, genetic feature selection (Additional File: Figure S5), model development, validation, and performance analysis are outlined in Figure 1. For benchmarking purposes, we also evaluated a number of standard ML models, including basic regression and tree-based methods such as Support Vector Machine (SVM) [26], Decision Tree [27], and Logistic Regression; probabilistic models such as Gaussian Naïve Bayes [28]; ensemble methods including Gradient Boosting [29], and XGBoost [30]; and the state-of-the-art tabular data processing model TabM [45]. For ML models requiring hyperparameter tuning, we utilized Ray Tune with the Asynchronous Successive Halving Algorithm (ASHA) based on distributed GPUs [31]. The optimization score was set to balanced accuracy defined as the average of sensitivity (true positive rate) and specificity (true negative rate) to try to prevent biased predictions. Hyperparameter search spaces and final chosen settings for baseline ML methods are presented in Additional File: Table S2. The operating

thresholds for ML models were determined on the training set using five-fold CV by selecting the probability threshold that maximized the balanced accuracy due to class imbalance issues, to account for class imbalance. For internal validation, we performed repeated five-fold cross-validation (Additional File: Figure S6), while external validation was conducted using bootstrapping methods. The relative importance of different patient measurements in determining predictive performance was computed using SHapely Additive exPlanation (SHAP) values [32]. Calibration plots were applied to evaluate the trustworthiness of the probabilistic predictions for each model [33]. For sample size estimation, we used the "pmsampsize" package in R [46], which yielded a minimum required sample size of 277 patients, including at least 66 events. For ML models, we employed the simulation-based approach proposed by van der Ploeg T et al. [44], where the minimum sample size required for each model is shown in the Additional File: Figure S1. Software packages used to perform the computations are also detailed in the (Additional File: Experimental environments).

Simulation of HR and RR measurement errors

To estimate the level of RR and HR measurement error likely to be encountered in clinical practice we utilized recordings of RR measurements from a convenience cohort of 25 adult patients who were admitted to hospital with acute illness, as reported by Drummond et al. [14], and HR error data in the supine position from the research by Kobayashi et al. [34]. For RR, the mean interquartile range (IQR) of

measurement error was 3.4, 3.0, and 2.5 breaths per minute corresponding to 30-, 60-, and 120-second counting windows, respectively. Specifically, we employed kernel density estimation (KDE) to model the error distribution in a flexible, data-driven manner [35], where the bandwidth and kernel function were selected through five-fold cross-validation by optimizing the log-likelihood score. For HR simulation, only means and percentiles by measurement duration/position were available [34]. We fit generalized lambda distributions (GLD) to match the reported percentiles and sampled perturbations (with physiologic truncation). Perturbed RR and HR values were generated through Monte Carlo sampling from the fitted probability function. The sampled values were then evaluated to exclude physiologically implausible observations. We defined physiologically implausible values using dataset-informed bounds with slightly permissive margins to avoid excluding plausible extremes. Specifically, for adults we reject and resample any simulated value with RR outside 5–60 bpm or HR outside 20–180 bpm. To verify that our Monte Carlo simulation approach produced realistic error data, we compared the simulated distributions against the RR error distributions from Drummond et al. [14] and HR error distributions from Kobayashi et al. [34]. Figure 2 describes how we matched the simulated errors to the original study [20, 34] and validates replication of the real clinical data.

Statistical Analysis

To assess whether simulated measurement errors reproduce the available ground-truth summaries, we analysed RR at the baseline level using two-sided paired sign tests (medians and IQRs) and trend diagnostics. For HR

where only aggregate summaries were available, we applied two-sided percentile-bootstrap tests for each statistic (mean, 10% percentile and 1% percentile) and an empirical joint test based on the bootstrap mean/covariance and Mahalanobis distance. To statistically measure the changes between baseline models and models including RR and HR error distributions, we performed a paired statistical comparison. Paired samples were aligned using bootstrap resampling indices and compared using either a paired t-test (if the distribution of differences was normal, as assessed by the Shapiro-Wilk test) or the Wilcoxon signed-rank test (non-parametric alternative for non-normal distributions). We computed both *P*-values and effect sizes, since *P*-values alone only indicate statistical significance but do not reflect the size of the change (Additional File: Table S3). This allowed us to assess both the statistical significance and the practical magnitude of differences between the baseline models and models including RR and HR error distributions. To measure whether the performance difference in discrimination ability (ROC curve) between models is significant, we used DeLong's test to statistically compare the ROC curves.

Continuous variables presented in Table 1 are reported as median and inter-quartile ranges and compared using non-parametric Mann-Whitney U test and Student's t-test based on the characteristics of the data. The Student's t-test was used for normally distributed data with equal variances, as determined by the Shapiro-Wilk test for normality and Levene's test for equal variances, or the Mann-Whitney U test was applied when applicable. Categorical variables were described by counts and

frequencies and compared using Fisher's exact test. To calculate the combined *P*-value for differences between the internal training and external validation cohorts, we applied Fisher's method. All tests were two-sided, and a *P*-value < 0.05 was considered statistically significant.

Results

Patient demographic and clinical characteristics

Some individual physiological parameters and clinical scores showed significant correlations with HFNC outcomes. As shown in Table 1, statistically significant differences between the HFNC failure and HFNC success groups in the internal training and/or external validation cohorts were observed for RR, FiO_2, and the ROX index and its variants (Student's t-test or Mann-Whitney U test, *P*-value < 0.05). In addition, the changes of key physiological variables such as PaO_2_diff, RR_diff, FiO_2_diff, and PaO_2/FiO_2_diff between timepoints T0 and T1 were also statistically significant, indicating their potential clinical relevance for outcome prediction. When comparing individual parameter or clinical indices between internal training and external validation cohorts, statistically significant inter-cohort differences were observed within both the success and failure subgroups, which suggest cohort-specific variability, making it challenging to determine a universally applicable cut-off value for any individual parameter or clinical index.

Monte Carlo simulations align with recorded RR and HR measurement errors

Based on the results of our Monte Carlo simulation with KDE (Figure 2), we can see that the simulated RR measurement errors assuming a 30-second counting time-window closely replicated the original data from [14]. Figure 2a also shows an example for a specific patient with a true RR of 17 breath/min, where the distribution of RR measurement errors follows a pattern similar to the distribution shown in the original study. It can also be observed in the original data that higher RR's produce a broader IQR of the RR measurement error, and our Monte Carlo simulations reproduce this trend. Across the 25 baselines, there was no evidence of systematic bias in medians ($P = 1.000$) or IQRs ($P = 0.890$), and no material drift of errors was found (Additional File: Figure S11). Similarly, for HR measurements, our simulation estimated the mean error of 1.8 beats/min for a 15-second counting window in supine position, with the probability of an absolute error exceeding 4 bpm being 5% (Figure 2b), as reported by Kobayashi et al. [36]. We compared simulated summaries to the reported values with two-sided bootstrap tests for each statistic; the reported values were typical under the simulator (mean: $P = 1.000$; 10% percentile: $P = 0.807$; 1% Percentile: $P = 0.506$). A joint empirical test based on the bootstrap mean/covariance of (mean, 10% percentile, 1% percentile) and Mahalanobis distance yielded $P = 0.701$, indicating overall consistency with the reported summaries.

HFNC outcome prediction performance using only non-invasive measurements

In the baseline evaluation without RR and HR measurement errors, the TabPFN model using only non-invasive measurements outperformed all

conventional ML models in predicting HFNC outcome in both internal and external datasets, as shown in Table 2. In addition, when compared with commonly used clinical indices, TabPFN demonstrated notable improvements in both discrimination and calibration across both internal and external validation cohorts, as shown in Table 3 and Figure 4. Among the clinical indices, the standard ROX index slightly outperformed its modified versions on average, with an accuracy of 0.67 (95% CI, 0.59-0.75), AUC of 0.76 (95% CI, 0.66-0.84), and Brier score of 0.196 during repeated five-fold cross-validation in the internal training cohort, and an accuracy of 0.72 (95% CI, 0.71-0.73), AUC of 0.80 (95% CI, 0.80-0.80), and Brier score of 0.183 in external validation. In comparison, the TabPFN model showed superior predictive capability, with an accuracy of 0.78 (95% CI, 0.72-0.87), AUC of 0.83 (95% CI, 0.75-0.89), and Brier score of 0.135 in internal cross-validation, and an accuracy of 0.77 (95% CI, 0.72-0.82), AUC of 0.84 (95% CI,0.81-0.87), and Brier score of 0.133 in external validation cohort.

To evaluate the impact of RR and HR measurement errors on predictive performance, we performed Monte Carlo sampling of these input variables to mimic the errors associated with short observation windows. The TabPFN model was more sensitive to these errors compared to clinical indices. As shown in Additional File: Figure S7 and Table S4, the difference between the baseline and perturbed ROC curves for TabPFN was statistically significant (DeLong's test, *P*-value < 0.01 without using AGB measurements). In contrast, the ROX indices and its variants showed no statistically significant difference in discrimination performance under the

same perturbation conditions (all *P*-values > 0.2). Furthermore, as shown in Figure 3a and 3c, TabPFN models using only non-invasive measurements demonstrated larger effect sizes across almost all performance metrics in both internal training and external validation cohorts. Therefore, despite TabPFN demonstrating superior predictive performance, discrimination, and calibration, more complex machine learning models such as TabPFN may be more sensitive to errors in RR and HR measurements.

HFNC outcome prediction performance using invasive measurements

When invasive arterial blood gas (ABG) measurements were also used, the TabPFN model showed improved predictive performance compared to when using only non-invasive measurements, with an accuracy of 0.80 (95% CI, 0.73-0.89), AUC of 0.85 (95% CI, 0.76-0.93), and Brier score of 0.127 during repeated five-fold cross-validation in the internal training cohort (Table 3, Figure 4). Similar improvements were observed in the external validation cohort, with an accuracy of 0.79 (95% CI, 0.74–0.85), AUC of 0.85 (95% CI, 0.81–0.89), and Brier score of 0.101. However, among the clinical indices, incorporating ABG measurements by replacing SpO_2 with PaO_2 (as in mROX and mROX-HR indices) did not yield better predictive outcomes, producing slightly lower discrimination performance (Figure 4) and lower performance accuracy in internal cross-validation compared to the original ROX and ROX-HR indices.

The integration of ABG measurements substantially improved the robustness of the TabPFN model to RR and HR measurement errors. In

the TabPFN model with ABG measurements, DeLong's test yielded a *P*-value > 0.05, indicating discriminative performance was not significantly affected by RR and HR perturbations. This contrasted with the non-invasive TabPFN model which showed a significantly higher sensitivity (*P*-value = 0.002). Moreover, the effect size of the TabPFN model incorporating ABG measurements was nearly half that of the TabPFN model using only non-invasive measurements (Figure 3). This can be explained by the interpretability analysis using SHAP plots (Figure 5 and Additional File: Figure S3), which shows that the PaO_2/FiO_2 ratio replaces RR as the most influential predictor when ABG measurements are available (Figure 5). For clinical indices incorporating PaO_2, DeLong's test yielded a *P*-value of 1.0 for both mROX_T1 and mROX-HR_T1, compared to *P*-values of 0.284 for ROX_T1 and 0.252 for ROX-HR_T1. In addition, as shown in heatmaps of effect size and *P*-value (Figure 3, Additional File: Table S2), there was no difference between mROX's baseline and perturbed performance (paired *t*-test *P*-value = 1.0, effect size=0) across nearly all evaluation metrics. mROX-HR_T1, with additional variability introduced by HR perturbations, exhibited slightly larger effect sizes than for mROX, though still smaller than those observed for ROX_T1 and ROX-HR_T1 without the use of PaO_2.

Effect of increasing the duration of RR and HR measurement windows

Increasing the durations of the counting windows for measuring RR and HR improved the stability of all models. Specifically, simulating a longer observation window (30 s for HR and 120 s for RR) with a resulting

reduction in the size of measurement errors significantly reduced their impact on the predictive accuracy of all models. This was reflected in a notable reduction in performance variability across both the internal and external validation cohorts, as demonstrated by smaller effect sizes when compared to simulations using shorter observation windows (15 seconds for HR and 30 seconds for RR), as shown in Figure 3c and 3d. Furthermore, when evaluating discriminative performance using DeLong's test, longer measurement durations yielded larger *P*-values across all models (see Additional File: Table S4), indicating a smaller difference between the baseline and perturbed ROC curves.

Discussion

In this study, we explored how errors in RR and HR measurements of a magnitude likely to occur in current clinical practice could affect the reliability and performance of models for predicting outcomes in patients undergoing HFNC therapy. The recently introduced TabPFN ML model was uniformly the most accurate predictor but was also more affected by measurement errors than simpler indices such as ROX. Augmenting patient datasets with ABG measurements further improved the performance of the TabPFN model and reduced its sensitivity to measurement errors. Interestingly, however, incorporating ABG measurements into the ROX index by replacing SpO_2 with PaO_2 (as in mROX and mROX-HR indices) did not yield better predictive performance (although it did reduce sensitivity to RR/HR measurement errors – Figure 3). This suggests that while using ABG measurements can improve

predictive accuracy in ML models such as TabPFN, their integration into clinical indices may not fully leverage the additional physiological information. PaO_2/FiO_2 ratio is the most influential predictor (Figure 5b) in the TabPFN model, but this model also makes use of multiple other features, uses both the level and the change in PaO_2/FiO_2 ratio, and learns nonlinear, conditional interactions (via attention) among ABG variables, respiratory rate, FiO_2, etc. This does not necessarily imply that replacing SpO_2 with PaO_2 in the ROX index will improve its predictive accuracy, however, because mROX employs fixed, shallow functional forms and typically does not include changes over time in PaO_2/FiO_2 ratio, making it unable to exploit ABG information when its effect depends on thresholds, trends, or interactions. Indeed we do not see any significant improvement in predictive performance when using the mROX index versus ROX in our data – as shown in Table 3, on the validation dataset, mROX_T1 has higher sensitivity but lower specificity, with similar accuracy and AUC to that achieved by ROX_T1.

Current RR and HR monitoring methods, which outside of ICU's predominantly rely on manual counting or intermittent recordings by time-pressured personnel in busy and noisy hospital environments, clearly have the potential to introduce measurement inaccuracies, particularly in acutely ill patients where minute-to-minute changes in breathing patterns and heartbeat sequence can be clinically significant [34, 36]. Aside from the use of longer counting windows, replacing manual counting with novel monitoring devices may offer more continuous and reliable measurements that could be of benefit in high-risk patients.

RR can be measured using an electrocardiogram, and several dedicated respiratory rate monitoring devices have also recently been developed to enhance the accuracy and reliability of RR measurements. *RespiraSense*, a novel piezoelectric-based respiratory rate monitor, is a motion-tolerant and continuous RR monitoring device that is now commercially available for clinical use [37, 38]. Studies have shown that such devices can operate from hospital admission through ambulation, ensuring uninterrupted tracking without restricting patient mobility [39]. Another promising device is the *Respeck* monitor, which is a non-invasive, tri-axial accelerometer-based wearable device [40, 41], developed to continuously monitor RR by sensing chest wall movement. This device also includes a ML-based signal processing system to further improve accuracy by filtering motion artifacts and noise, allowing the better identification of valid breaths, particularly in active or acutely ill patients. Finally, a novel method for RR monitoring that integrates a small pressure transducer into the HFNC system itself was recently tested in healthy volunteers [42]. Unlike external motion sensors, this method leverages the potential capability of the HFNC system to detect actual inspiratory and expiratory events with high fidelity. The pressure variations during breathing may not only provide a precise RR measurement but also serve as a potential marker of respiratory effort.

A number of novel wearable optical HR sensors are also now available, ranging from consumer- to research-grade systems. A recent study comparing the effects of different skin tones, motion artifacts, and signal

crossover found that devices with higher cost, a more recent release data, and a larger market, were more accurate [43].

The main clinical implications of this study are the following. The predictive performance of currently available and easy to calculate indices (ROX and its variants) appears to be quite insensitive to measurement errors in RR and HR, increasing confidence in their applicability in realistic clinical scenarios. Our results also suggest that recently proposed ML models could deliver higher predictive performance, but for their potential to be fully realised, prospective studies using these approaches should try to minimise RR/HR measurement errors (by using longer manual counting windows or novel monitoring devices) and/or incorporate additional measurements such as ABG's. More generally, as we enter the era of data-driven healthcare, more attention should be paid to rigorously quantifying the effects of inevitable errors in data on the predictions that are derived from them.

This study has some limitations. The total number of patients available in our datasets is lower than what would typically be used for the training and validation of standard ML models. However, the TabPFN model deployed here has been specifically developed to work with small-to-medium-sized datasets, while the granularity of the measurements available in the analysed datasets is extremely high, and strengthens the performance of the model. Also, the data on the size of RR and HR measurement errors likely to occur in clinical practice that was used for our simulations comes from single site studies with limited sample size, and therefore the full potential spectrum of possible measurement errors

may not be completely captured. The lack of access to patient-level raw data for the external sources [14, 34], necessitated the use of summary-statistic-based modelling - we mitigated this with quantile-matched non-parametric estimation. Finally, due to the unavailability of specific data, potential errors in the measurement of SpO_2 and FiO_2 which may also affect prediction accuracy could not be included in our analysis.

Conclusions

This multi-centre study has performed the first systematic analysis of the effect of heart and respiratory rate measurement errors on the ability to predict the outcome of HFNC therapy soon after initiation. Clinically realistic levels and distributions of errors in both variables were derived from previous *in vivo* studies and simulated using rigorous statistical methods.

The ROX index (and its variants) were found to be relatively insensitive to these errors. Machine learning models produced higher predictive performance but were also more affected by measurement errors. The recently proposed TabPFN in-context learning model was able to accurately and robustly predict the outcome of HFNC, using only small-to-medium-sized datasets made up of routinely available measurements made early in the patients' treatment, while accounting for potential errors in measured respiratory and heart rates. Reducing these errors by using longer counting windows or dedicated RR and HR measurement sensors, and/or incorporating arterial blood gas measurements into their training data will help maximise the predictive accuracy of ML models, allowing

earlier identification of patients who are at risk of failing HFNC so that their treatment can be adjusted or escalated in a timely fashion.

Table 1. Characteristics of patients in the internal training and external validation cohorts. Data are presented as median (interquartile range, 25%-75%) for continuous values unless otherwise specified. *P* for difference between HFNC failure vs. success. *P** for difference between the internal training cohort and the external validation cohort.

Variables	Internal Training Cohort		*P*	External Validation Cohort		*P*	*P**
	HFNC Failure (N=140)	HFNC Success (N=456)		HFNC Failure (N=85)	HFNC Success (N=156)		
Age, y	58 (44, 70)	63 (52, 76)	0.004	68 (65, 79)	68 (66, 76)	0.102	<0.001
Baseline measurements (within 6 hours prior to or at the initiation of HFNC) (T0 time)							
$PaCO_2$ (T0), mmHg	46.7 (37.5, 54.5)	44.0 (37.3, 47.8)	0.958	32.1 (30.4, 34.2)	33.4 (31.6, 34.5)	0.084	<0.001
PaO_2 (T0), mmHg	60.1 (40.0, 83.0)	82.3 (51.0, 98.8)	0.242	64.1 (59.6, 70.2)	66.5 (60.0, 73.8)	0.031	<0.001
FiO_2 (T0), %	57 (51, 66)	46 (33, 66)	<0.001	56 (50, 60)	49 (40, 60)	0.077	<0.001
PaO_2/FiO_2 (T0), mmHg	143 (77, 183)	130 (88, 178)	<0.001	119 (96, 140)	143 (118, 168)	0.006	<0.001
SpO_2 (T0), %	94.0 (92.0, 96.0)	92.8 (91.0, 96.0)	0.003	92.0 (90.0, 95.0)	92.8 (91.0, 95.0)	0.035	0.021
RR (T0), bpm	27 (23, 30)	25 (21, 28)	0.001	30 (25, 36)	27 (24, 28)	0.002	0.025
HR (T0), bpm	91 (77, 102)	87 (75, 99)	0.104	95 (79, 104)	94 (86, 102)	0.179	<0.001
ROX (T0)	5.6 (3.8, 7.0)	8.0 (5.4, 9.8)	<0.001	6.0 (4.6, 6.9)	7.8 (5.9, 9.1)	<0.001	0.328
ROX_HR (T0)	8.2 (5.2, 9.5)	12.0 (7.4, 15.2)	<0.001	6.6 (4.7, 7.7)	8.6 (6.4, 9.9)	<0.001	<0.001
mROX (T0)	7.0 (3.4, 10.6)	7.6 (4.1, 10.8)	<0.001	4.2 (3.1, 5.2)	5.6 (4.0, 6.8)	<0.001	<0.001
mROX_HR (T0)	14.3 (5.7, 16.2)	12.1 (7.6, 14.3)	<0.001	4.6 (3.0, 5.5)	6.2 (4.2, 7.7)	<0.001	<0.001
1-2 h after HFNC initiation (T1 time)							
$PaCO_2$ (T1), mmHg	43.2 (35.0, 49.5)	42.1 (35.5, 44.0)	0.939	32.6 (30.8, 35.0)	35.4 (33.2, 36.9)	<0.001	0.048

PaO_2 (T1), mmHg	93.0 (40.0,135.0)	94.4(69.3,102.0)	0.014	67.2 (60.4, 72.7)	66.0 (61.2, 70.0)	0.344	<0.001
FiO2 (T1), %	71 (60, 89)	57 (42, 63)	<0.001	63 (55, 70)	45 (35, 51)	<0.001	0.031
PaO_2/FiO_2 (T1), mmHg	131 (68, 173)	145 (95, 182)	<0.001	115 (97, 131)	158 (130, 192)	<0.001	0.166
SpO2 (T1), %	93 (91, 95)	94 (93, 96)	<0.001	94 (93, 95)	94 (92, 95)	0.493	0.865
RR (T1) bpm	27 (23, 30)	24 (20, 27)	<0.001	29 (26, 36)	21 (19, 24)	<0.001	0.002
HR (T1), bpm	91 (77, 102)	85 (71, 97)	0.005	95 (86, 103)	90 (80, 100)	0.004	<0.001
ROX (T1)	5.6 (4.1, 6.6)	8.9 (5.7, 10.7)	<0.001	5.6 (4.9, 6.3)	10.7 (8.3, 13.1)	<0.001	<0.001
ROX_HR (T1)	6.3 (4.4, 7.6)	10.4 (6.6, 12.5)	<0.001	6.1 (4.8, 7.3)	12.4 (9.1, 14.8)	<0.001	<0.001
mROX (T1)	6.3 (3.1, 8.1)	8.4 (4.7, 10.5)	<0.001	4.0 (3.3, 4.7)	7.5 (5.6, 8.8)	<0.001	0.967
mROX_HR (T1)	7.2 (2.8, 8.7)	10.4 (6.0, 13.1)	<0.001	4.4 (3.3, 5.4)	8.7 (6.2, 10.5)	<0.001	0.180
Change in measurements between two time points							
$PaCO_2$_diff, mmHg	-3.6 (-4.0, 0.5)	-1.9 (-4.0, 4.3)	0.808	0.4 (-1.0, 2.3)	2.1 (0.0, 3.9)	0.001	<0.001
PaO_2_diff, mmHg	32.9 (-1.0, 48.5)	12.1(-19.8, 34.5)	<0.001	3.1 (-4.8, 11.7)	-0.5 (-8.7, 6.3)	<0.001	0.181
FiO_2_diff	14.4 (0.0, 27.3)	10.4 (-0.8, 20.0)	<0.001	7.4 (-2.2, 15.0)	-4.8 (-10.0, 0.0)	<0.001	<0.001
PaO_2/FiO_2_diff, mmHg	-11 (-93, 50)	-15 (-25, 53)	0.015	-4 (-16, 12)	14 (2, 25)	<0.001	<0.001
SpO_2_diff	0.1 (-2.0, 3.0)	0.1 (-1.0, 2.0)	0.666	1.7 (0.0, 4.0)	1.0 (-1.0, 2.0)	0.006	0.009
RR_diff, bpm	0 (-2, 3)	-1 (-3, 1)	<0.001	-2 (-4, 1)	-5 (-7, -2)	<0.001	<0.001
HR _diff, bpm	0 (-4, 4)	-3 (-8, 3)	0.084	0 (-4, 8)	-4 (-8, 2)	0.026	0.102

Table 1. Performance comparisons of different machine learning models. "Training" refers to the results from 100× repeated five-fold cross-validation conducted on the RENOVATE dataset, while "validation" corresponds to external validation on the combined Modena, MIMIC-IV, and eICU datasets. Metrics were obtained through 200 times bootstrapping and are reported as the mean with 95% confidence intervals.

Model/Indices	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Non-ABG measurements						
TabPFN (non-ABG) (Training)	0.78 [0.72-0.87]	0.83 [0.70-0.96]	0.77 [0.69-0.87]	0.63 [0.52-0.74]	0.84 [0.79-0.91]	0.83 [0.75-0.89]
TabPFN (non-ABG) (Validation)	0.77 [0.72-0.82]	0.85 [0.76-0.93]	0.75 [0.64-0.84]	0.69 [0.63-0.74]	0.87 [0.83-0.92]	0.84 [0.81-0.87]
SVM (non-ABG) (Training)	0.74 [0.64-0.83]	0.74 [0.55-0.93]	0.74 [0.60-0.84]	0.54 [0.44-0.67]	0.81 [0.76-0.87]	0.80 [0.73-0.87]
SVM (non-ABG) (Validation)	0.75 [0.69-0.79]	0.70 [0.63-0.82]	0.79 [0.71-0.85]	0.68 [0.60-0.75]	0.82 [0.77-0.86]	0.81 [0.75-0.86]
Logistic (Non-ABG) (Training)	0.69 [0.63-0.75]	0.68 [0.52-0.85]	0.70 [0.63-0.77]	0.53 [0.35-0.76]	0.84 [0.81-0.88]	0.77 [0.67-0.86]
Logistic (Non-ABG) (Validation)	0.73 [0.67-0.78]	0.69 [0.63-0.77]	0.76 [0.68-0.82]	0.63 [0.55-0.70]	0.82 [0.77-0.86]	0.79 [0.73-0.85]
DecisionTree (Non-ABG) (Training)	0.72 [0.63-0.80]	0.64 [0.44-0.82]	0.75 [0.65-0.84]	0.41 [0.37-0.53]	0.83 [0.79-0.89]	0.76 [0.67-0.79]
DecisionTree (Non-ABG) (Validation)	0.71 [0.67-0.75]	0.61 [0.49-0.73]	0.77 [0.71-0.82]	0.65 [0.56-0.75]	0.73 [0.70-0.77]	0.76 [0.70-0.83]
XGBoost (Non-ABG) (Training)	0.71 [0.62-0.80]	0.74 [0.56-0.92]	0.71 [0.61-0.81]	0.51 [0.41-0.63]	0.80 [0.74-0.86]	0.79 [0.69-0.87]
XGBoost Non-ABG) (Validation)	0.72 [0.69-0.75]	0.75 [0.66-0.82]	0.72 [0.70-0.74]	0.62 [0.53-0.72]	0.71 [0.68-0.73]	0.82 [0.79-0.84]
GradientBoost (Non-ABG) (Training)	0.74 [0.67-0.81]	0.76 [0.57-0.95]	0.75 [0.67-0.83]	0.55 [0.44-0.70]	0.81 [0.76-0.84]	0.78 [0.66-0.89]
GradientBoost (Non-ABG) (Validation)	0.71 [0.67-0.75]	0.71 [0.61-0.81]	0.73 [0.70-0.76]	0.73 [0.61-0.84]	0.71 [0.68-0.73]	0.81 [0.77-0.84]
GaussianNB (Non-ABG) (Training)	0.68 [0.61-0.74]	0.63 [0.40-0.85]	0.73 [0.65-0.80]	0.50 [0.35-0.68]	0.84 [0.80-0.90]	0.77 [0.67-0.86]
GaussianNB (Non-ABG) (Validation)	0.70 [0.65-0.74]	0.65 [0.56-0.78]	0.75 [0.71-0.78]	0.67 [0.58-0.78]	0. 74 [0.70-0.77]	0.79 [0.76-0.82]
TabM (Non-ABG) (Training)	0.76 [0.68-0.88]	0.74 [0.58-0.92]	0.79 [0.68-0.89]	0.61 [0.49-0.73]	0.83 [0.76-0.91]	0.80 [0.71-0.89]

TabM (Non-ABG) (Validation)	0.74 [0.66-0.83]	0.81 [0.72-0.91]	0.77 [0.63-0.89]	0.65 [0.57-0.76]	0.84 [0.80-0.93]	0.80 [0.72-0.89]
ABG measurements available						
TabPFN (ABG incl.) (Training)	0.80 [0.73-0.89]	0.84 [0.78-0.90]	0.78 [0.69-0.87]	0.62 [0.51-0.77]	0.85 [0.81-0.88]	0.85 [0.78-0.93]
TabPFN (ABG incl.) (Validation)	0.79 [0.74-0.85]	0.78 [0.68-0.90]	0.77 [0.65-0.86]	0.63 [0.55-0.71]	0.87 [0.82-0.92]	0.85 [0.81-0.89]
SVM (ABG incl.) (Training)	0.76 [0.69-0.81]	0.80 [0.65-0.95]	0.76 [0.66-0.84]	0.51 [0.43-0.60]	0.93 [0.89-0.98]	0.81 [0.72-0.89]
SVM (ABG incl.) (Validation)	0.75 [0.71-0.79]	0.77 [0.74-0.88]	0.75 [0.69-0.81]	0.67 [0.61-0.72]	0.84 [0.81-0.88]	0.80 [0.76-0.83]
Logistic (ABG incl.) (Training)	0.72 [0.65-0.79]	0.73 [0.55-0.93]	0.73 [0.68-0.80]	0.54 [0.40-0.71]	0.81 [0.77-0.86]	0.79 [0.71-0.87]
Logistic (ABG incl.) (Validation)	0.73 [0.72-0.74]	0.73 [0.66-0.77]	0.72 [0.67-0.79]	0.64 [0.60-0.71]	0.77 [0.74-0.79]	0.81 [0.78-0.84]
DecisionTree (ABG incl.) (Training)	0.67 [0.57-0.76]	0.74 [0.57-0.89]	0.64 [0.53-0.77]	0.46 [0.39-0.56]	0.88 [0.81-0.92]	0.76 [0.69-0.84]
DecisionTree (ABG incl.) (Validation)	0.71 [0.69-0.73]	0.86 [0.81-0.90]	0.58 [0.51-0.65]	0.62 [0.60-0.65]	0.84 [0.81-0.87]	0.78 [0.75-0.82]
XGBoost (ABG incl.) (Training)	0.76 [0.70-0.82]	0.72 [0.53-0.90]	0.78 [0.60-0.84]	0.56 [0.38-0.77]	0.84 [0.79-0.89]	0.79 [0.72-0.89]
XGBoost Non-ABG) (Validation)	0.73 [0.70-0.77]	0.69 [0.60-0.68]	0.77 [0.75-0.78]	0.74 [0.64-0.82]	0.76 [0.73-0.79]	0.79 [0.77-0.91]
GradientBoost (ABG incl.) (Training)	0.75 [0.69-0.84]	0.74 [0.55-0.97]	0.79 [0.66-0.85]	0.57 [0.36-0.72]	0.85 [0.74-0.87]	0.79 [0.67-0.91]
GradientBoost (ABG incl.) (Validation)	0.77 [0.71-0.83]	0.73 [0.65-0.84]	0.75 [0.58-0.87]	0.56 [0.40-0.73]	0.83 [0.75-0.83]	0.82 [0.76-0.88]
GaussianNB (ABG incl.) (Training)	0.68 [0.56-0.75]	0.63 [0.40-0.85]	0.73 [0.65-0.80]	0.50 [0.35-0.68]	0.84 [0.80-0.90]	0.77 [0.67-0.86]
GaussianNB (ABG incl.) (Validation)	0.71 [0.69-0.73]	0.76 [0.57-0.88]	0.57 [0.49-0.66]	0.57 [0.61-0.67]	0.82 [0.80-0.88]	0.78 [0.75-0.82]
TabM (ABG incl.) (Training)	0.77 [0.67-0.83]	0.78 [0.64-0.95]	0.77 [0.63-0.87]	0.53 [0.41-0.62]	0.80 [0.71-0.93]	0.82 [0.71-0.90]
TabM (ABG incl.) (Validation)	0.74 [0.71-0.79]	0.76 [0.73-0.89]	0.76 [0.63-0.89]	0.68 [0.60-0.74]	0.83 [0.77-0.90]	0.80 [0.73-0.86]

Table 2. External validation of best-performing machine learning models and clinical indices at baseline and with simulated errors in respiratory and heart rates. "Training" refers to the results from 100× repeated five-fold cross-validation conducted on the RENOVATE dataset, while "validation" corresponds to external validation on the combined Modena, MIMIC-IV, and eICU datasets. Metrics were obtained through 200 times bootstrapping and are reported as the mean with 95% confidence intervals. * indicates that the data includes simulated heart rate errors corresponding to a counting window of 15 seconds in the supine position and simulated respiratory rate errors corresponding to a counting window of 30 seconds.

Model/Indices	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
TabPFN (non-ABG) (Training)	0.78 [0.72-0.87]	0.83 [0.70-0.96]	0.77 [0.69-0.87]	0.63 [0.52-0.74]	0.84 [0.79-0.91]	0.83 [0.75-0.89]
TabPFN (non-ABG)* (Training)	0.76 [0.70-0.85]	0.80 [0.66-1.00]	0.75 [0.68-0.87]	0.61 [0.50-0.74]	0.84 [0.78-0.95]	0.81 [0.72-0.89]
TabPFN (non-ABG) (Validation)	0.77 [0.72-0.82]	0.85 [0.76-0.93]	0.75 [0.64-0.84]	0.69 [0.63-0.74]	0.87 [0.83-0.92]	0.84 [0.81-0.87]
TabPFN (non-ABG)* (Validation)	0.79 [0.73-0.84]	0.88 [0.78-0.96]	0.77 [0.64-0.88]	0.70 [0.63-0.78]	0.90 [0.84-0.95]	0.86 [0.82-0.89]
TabPFN (ABG incl.) (Training)	0.80 [0.73-0.89]	0.84 [0.78-0.90]	0.78 [0.69-0.87]	0.62 [0.51-0.77]	0.85 [0.81-0.88]	0.85 [0.76-0.93]
TabPFN (ABG incl.)* (Training)	0.80 [0.71-0.89]	0.84 [0.73-0.92]	0.78 [0.68-0.87]	0.62 [0.50-0.77]	0.85 [0.80-0.90]	0.85 [0.75-0.94]
TabPFN (ABG incl.) (Validation)	0.79 [0.74-0.85]	0.78 [0.68-0.90]	0.77 [0.65-0.86]	0.63 [0.55-0.71]	0.87 [0.82-0.92]	0.85 [0.81-0.89]
TabPFN (ABG incl.)* (Validation)	0.81 [0.75-0.86]	0.75 [0.61-0.87]	0.81[0.70-0.90]	0.66 [0.58-0.74]	0.86 [0.80-0.91]	0.86 [0.81-0.90]
ROX_T1 (Training)	0.67 [0.59-0.75]	0.76 [0.59-0.91]	0.65 [0.54-0.75]	0.37 [0.29-0.45]	0.91 [0.82-0.96]	0.76 [0.66-0.84]
ROX_T1* (Training)	0.65 [0.56-0.74]	0.77 [0.59-0.91]	0.61 [0.50-0.73]	0.35 [0.27-0.43]	0.91 [0.85-0.96]	0.74 [0.63-0.83]
ROX_T1 (Validation)	0.72 [0.71-0.73]	0.73 [0.69-0.76]	0.71 [0.68-0.76]	0.66 [0.65-0.69]	0.77 [0.75-0.78]	0.80 [0.80-0.80]
ROX_T1* (Validation)	0.71 [0.68-0.75]	0.71 [0.63-0.79]	0.72 [0.65-0.78]	0.67 [0.62-0.71]	0.76 [0.71-0.81]	0.78 [0.75-0.80]
mROX_T1 (Training)	0.65 [0.57-0.74]	0.75 [0.59-0.90]	0.62 [0.52-0.71]	0.38 [0.33-0.42]	0.88 [0.82-0.94]	0.73 [0.62-0.83]
mROX_T1* (Training)	0.66 [0.59-0.76]	0.75 [0.55-0.91]	0.62 [0.52-0.71]	0.38 [0.34-0.44]	0.88 [0.81-0.94]	0.73 [0.64-0.85]

mROX_T1 (Validation)	0.72 [0.71-0.72]	0.84 [0.81-0.85]	0.62 [0.59-0.64]	0.64 [0.62-0.64]	0.83 [0.81-0.84]	0.80 [0.80-0.80]
mROX_T1* (Validation)	0.72 [0.71-0.76]	0.84 [0.81-0.85]	0.62 [0.59-0.64]	0.64 [0.62-0.64]	0.83 [0.81-0.84]	0.80 [0.80-0.80]
ROX_HR_T1 (Training)	0.67 [0.59-0.75]	0.76 [0.59-0.91]	0.64 [0.54-0.74]	0.36 [0.29-0.45]	0.91 [0.85-0.96]	0.77 [0.68-0.85]
ROX_HR_T1* (Training)	0.64 [0.55-0.73]	0.77 [0.59-0.91]	0.61 [0.49-0.72]	0.35 [0.27-0.43]	0.91 [0.84-0.96]	0.75 [0.66-0.84]
ROX_HR_T1 (Validation)	0.72 [0.71-0.74]	0.75 [0.68-0.81]	0.70 [0.69-0.73]	0.66 [0.65-0.68]	0.78 [0.74-0.82]	0.80 [0.80-0.80]
ROX_HR_T1* (Validation)	0.71 [0.67-0.75]	0.74 [0.66-0.82]	0.69 [0.62-0.76]	0.66 [0.61-0.70]	0.77 [0.72-0.82]	0.79 [0.77-0.81]
mROX_HR_T1 (Training)	0.64 [0.58-0.74]	0.74 [0.59-0.90]	0.60 [0.50-0.70]	0.38 [0.32-0.36]	0.89 [0.84-0.96]	0.74 [0.64-0.84]
mROX_HR_T1* (Training)	0.63 [0.56-0.73]	0.75 [0.59-0.91]	0.58 [0.48-0.69]	0.38 [0.31-0.45]	0.90 [0.83-0.96]	0.74 [0.63-0.83]
mROX_HR_T1 (Validation)	0.71 [0.69-0.73]	0.86 [0.81-0.90]	0.58 [0.51-0.65]	0.62 [0.60-0.65]	0.84 [0.81-0.87]	0.78 [0.78-0.78]
mROX_HR_T1* (Validation)	0.70 [0.69-0.73]	0.87 [0.81-0.90]	0.57 [0.50-0.65]	0.62 [0.59-0.65]	0.85 [0.81-0.88]	0.78 [0.78-0.78]

Figure 1. Overview of the TabPFN model development, validation, and effect analysis process

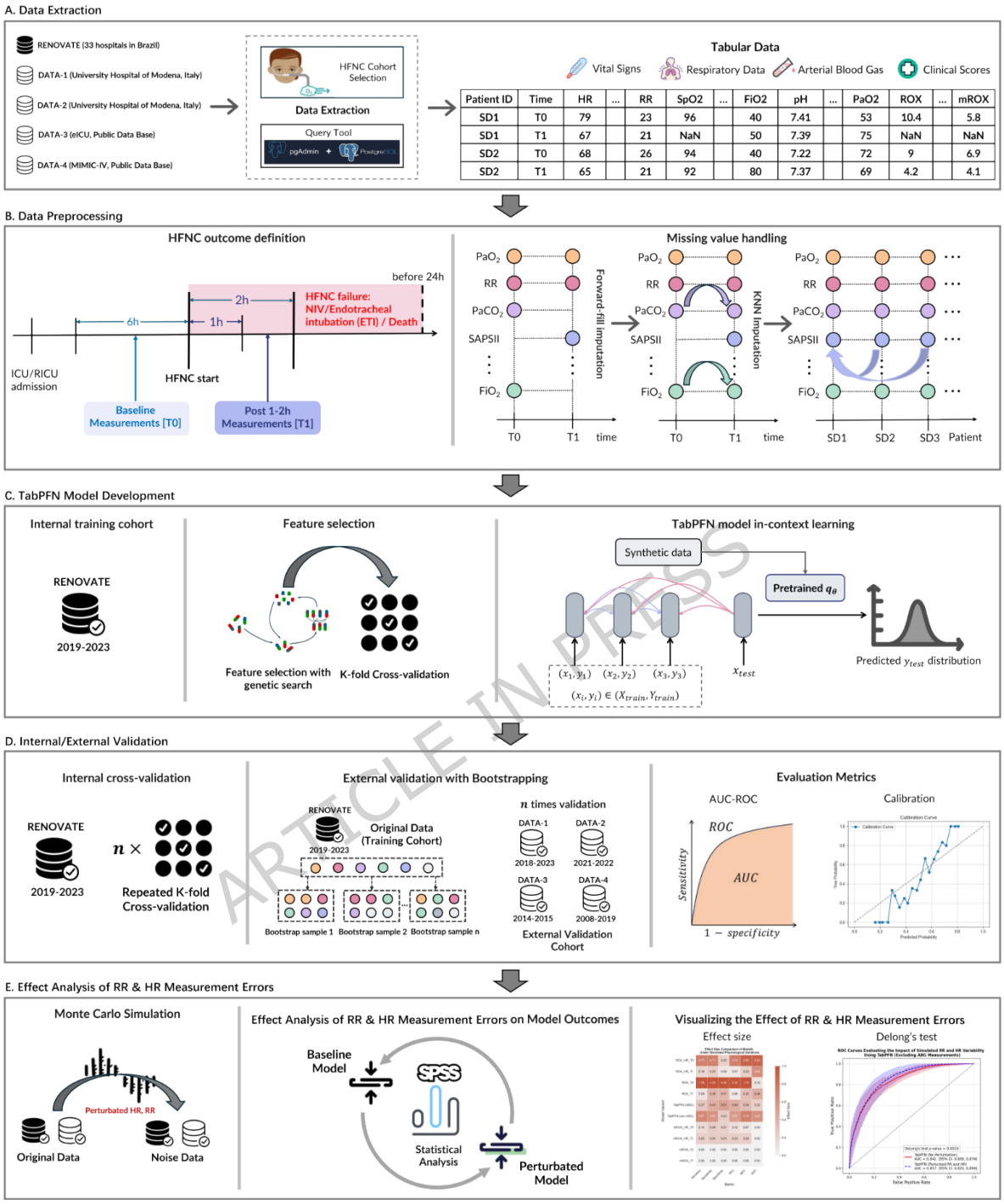


Figure 2. Monte Carlo Simulation of RR and HR Measurement Errors. (a) RR measurements from 25 subjects using a 30-second counting window [14]. The y-axis represents the median and interquartile range of measured RR, plotted against the true RR on the x-axis for each subject. Green dots indicate the original measured data, while red dots represent Monte Carlo-simulated RR values based on kernel density estimation. Left panel: An example patient's simulated distribution of RR measurements is shown, with the true RR indicated by a black dot (●). (b) Bland–Altman plot showing the agreement between measured and true heart rate (HR) during a 15-second counting interval in the supine position [34]. Each point represents the error (measured – true HR) plotted against the mean of the two measurements. The fitted mean unsigned error is 1.77 bpm (green dashed line), with the 95% confidence interval shaded. Percentile values of absolute error are 4 bpm at the 10th and 5th percentiles, and 6 bpm at the 1st percentile. According to study [34], "5th percentile = 4" indicates that only 5% of the errors are expected to exceed 4 bpm.

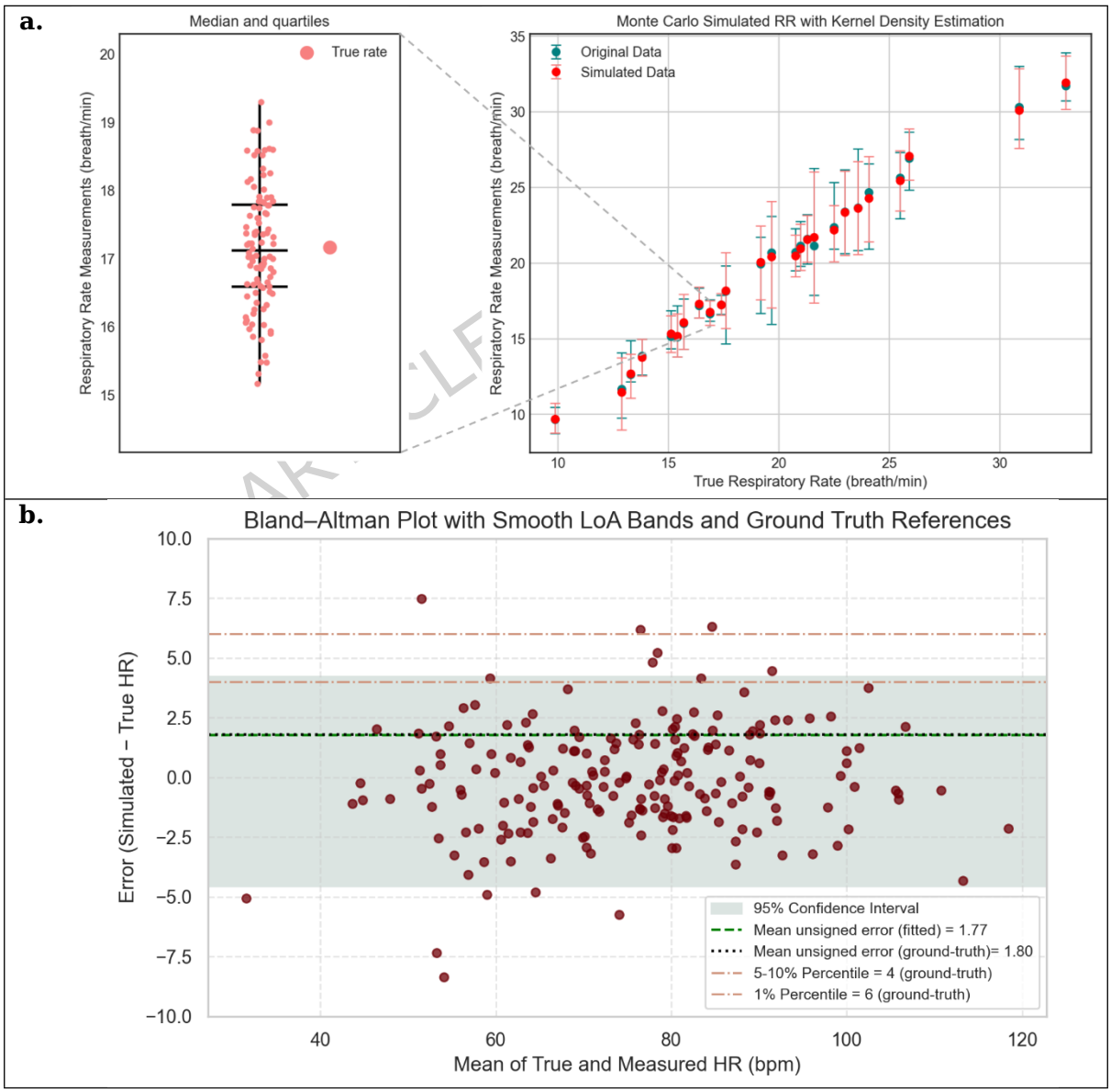


Figure 3. Effect Size Heatmaps: Baseline Model Performance vs. Simulated RR and HR Measurement Errors Across Varying Measurement Durations. Effect sizes illustrate the magnitude of performance changes due to simulated HR and RR measurement errors. Larger values indicate greater sensitivity to perturbations. (a, c) Heatmaps comparing baseline and perturbed models with HR errors arising from a 15 second counting window and RR errors arising from a 30 seconds counting window. (b, d) Comparisons for smaller errors due to longer counting windows: HR (30 seconds) and RR (120 seconds). Panels (a, b) internal training cohort using repeated five-fold cross-validation with Monte Carlo simulation; panels (c, d) external validation cohort using bootstrapping and Monte Carlo simulation.

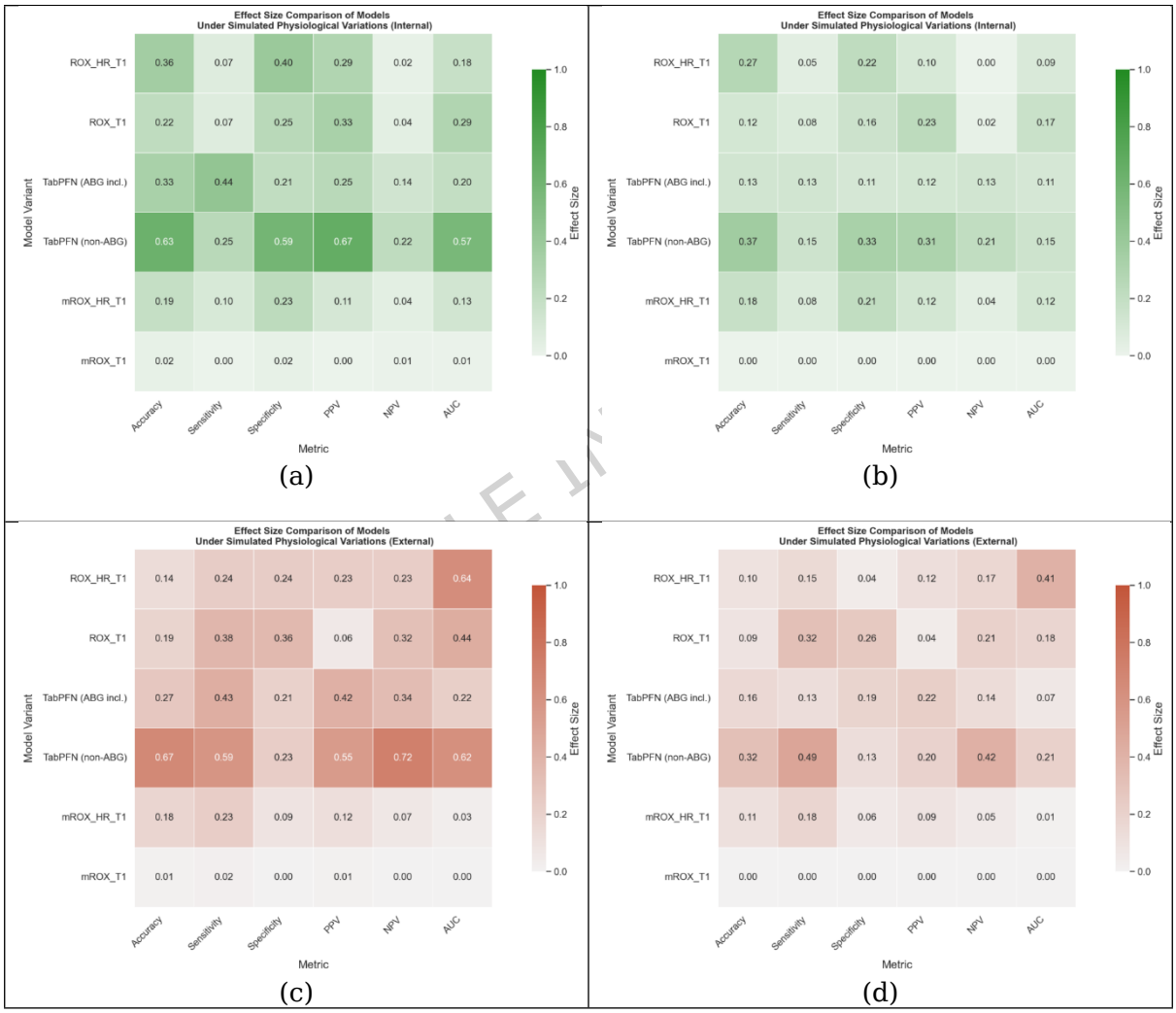


Figure 4. Discrimination and Calibration Performance of TabPFN and Clinical Indices on both Internal and External Validation Cohort. (a, c) The Receiver Operating Characteristic (ROC) curve comparing predictive performance (b, d) Calibration curves for the model's predictions - a perfectly calibrated model, where the predicted probability precisely matches the observed frequencies, would follow the dashed diagonal line. The upper panels show results from the internal training cohort, obtained through 100× repeated five-fold cross-validation, with values reported as the mean across all repetitions. The lower panels present results from the external validation cohort, calculated using 200× bootstrapping, with values reported as the mean across all iterations.

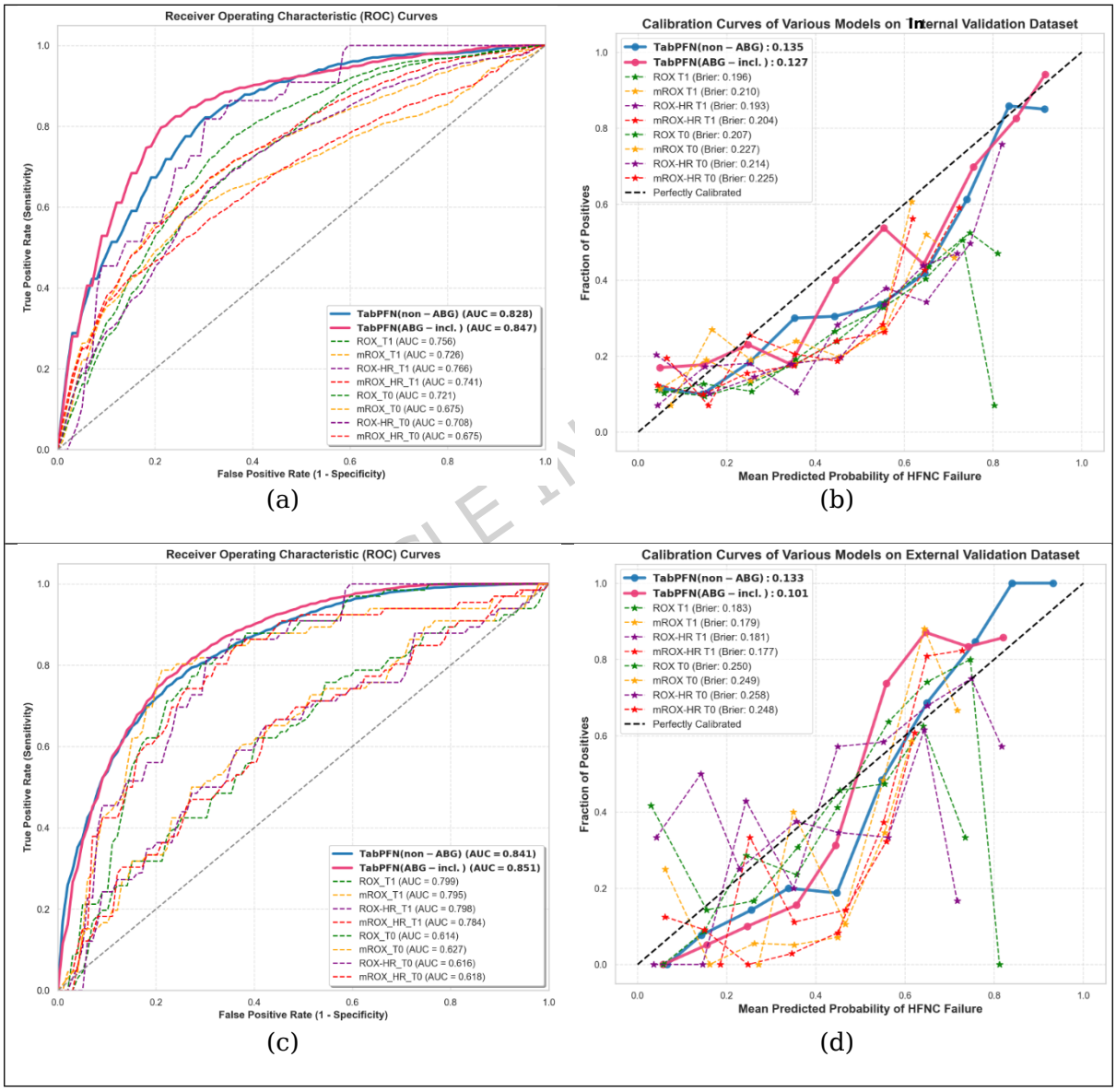
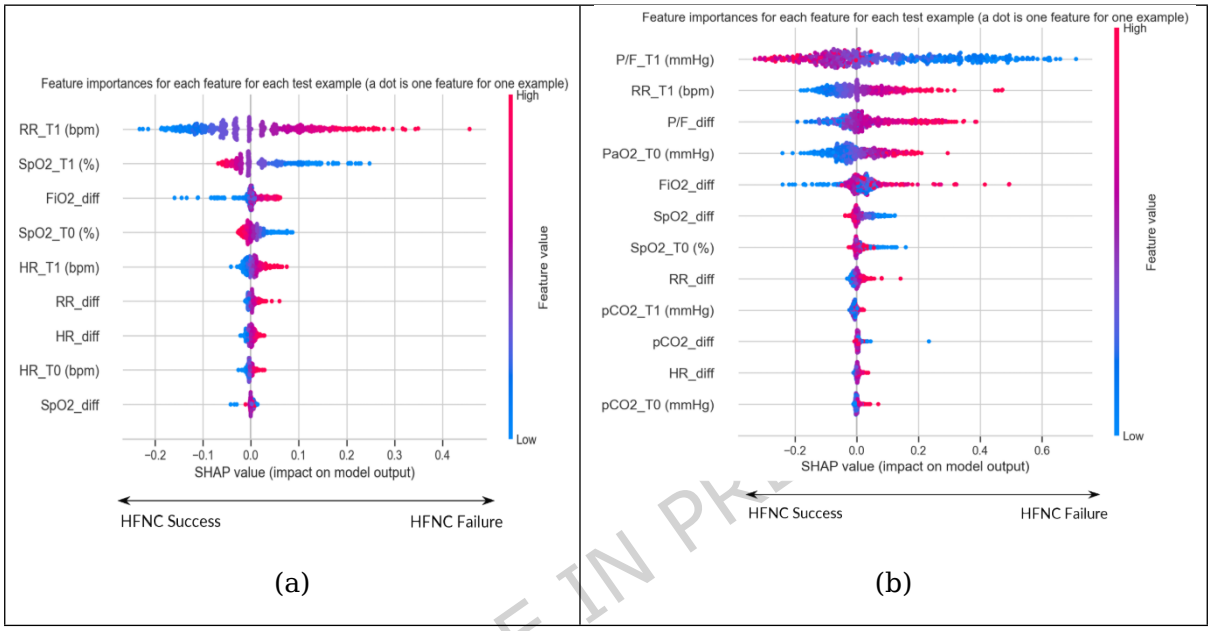


Figure 5. SHAP summary plot for the TabFPN model. Horizontal axis: The impact of each feature on the model's prediction. Positive SHAP values indicate that the feature contributes to predicting a HFNC failure, while negative SHAP values indicate a contribution to predicting HFNC success. Vertical axis: The list of features used for making a prediction, ordered based on their importance, with the most important features at the top. Dots: Each dot represents a single patient in the dataset. (a) TabPFN using only non-invasive measurements. (b) TabPFN using invasive measurements.



Availability of data and material

Patient data for internal validation and code used for the machine learning model are freely available on request by bona fide researchers for specified scientific purposes from the corresponding author. Data for external validation are publicly accessible, including deidentified patient data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) v2.2 database at https://physionet.org/content/mimiciv/2.2/ and from the eICU Collaborative Research Database (eICU) at https://eicu-crd.mit.edu/. The RENOVATE dataset used for model development is available upon request to bona fide researchers for specific scientific purposes, subject to approval by the RENOVATE investigators. The repository containing all code for reproducibility can be found at https://github.com/BioTechDog/Stress-Testing-HFNC-Prediction-Models.git.

Abbreviations

HFNC: High-flow nasal cannula

ML: Machine learning

AHRF: Acute hypoxemic respiratory failure

ICU: Intensive care unit

ML: Machine learning

SHAP: Shapely additive explanation

ROX: Ratio of oxygen saturation as measured by pulse oximetry/oxygen fraction index

RR: Respiratory rate

HR: Heart rate

SpO_2: Saturation of pulse oxygen

PaO_2/FiO_2: The ratio of partial pressure of oxygen in arterial blood to the fraction of inspiratory oxygen concentration

ABG: Arterial blood gas

Ethics statement

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Competing interests

The authors declare no competing interests.

Author contributions

HY: Conceptualization, Formal analysis, Methodology, Software, Writing – original draft. SS: Conceptualization, Methodology, Writing – review & editing. RT: Data curation, Methodology, Writing – review & editing. JGL: Methodology, Writing – review & editing. QZ: Writing – review & editing. AE: Writing – review & editing. LL: Data curation, Writing – review & editing. LK-D: Data curation, Writing – review & editing. IM: Data curation, Writing – review & editing. AC: Data curation, Writing – review & editing. EC: Data curation, Methodology, Writing – review & editing. DGB: Conceptualization, Supervision, Writing – original draft.

Funding

This work was supported by the UKRI Engineering and Physical Sciences Research Council (Ref. EP/W000490/1) and The Royal Academy of Engineering (Ref. RF2122-21–258).

References

1. Smith MEB, Chiovaro JC, Neil MO, et al. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Ann Am Thorac Soc*. (2014)11(9):1454–65.
2. Frat JP, Thille AW, Mercat A, Girault C, Ragot S, Perbet S, et al. High-flow oxygen through nasal cannula in acute hypoxemic respiratory failure. *N Engl J Med*. (2015) 372(23):2185–96.
3. Frat JP, Ragot S, Girault C, Perbet S, Prat G, Boulain T, et al. REVA network effect of non-invasive oxygenation strategies in immunocompromised patients with severe acute respiratory failure: a post-hoc analysis of a randomised trial. *Lancet Respir Med*. (2016) 4(8):646–52.
4. Oczkowski S, Ergan B, Bos L, et al. ERS clinical practice guidelines: high-flow nasal cannula in acute respiratory failure. *Eur Respir J*. (2022) 59(4):2101574.
5. Roca O, Caralt B, Messika J, Samper M, Sztrymf B, Hernandez G, et al. An index combining respiratory rate and oxygenation to predict outcome of nasal high-flow therapy. *Am J Respir Crit Care Med*. (2019)199(11):1368–76.
6. Goh KJ, Chai HZ, Ong TH, Sewa DW, Phua GC, Tan QL. Early prediction of high-flow nasal cannula therapy outcomes using a modified ROX index incorporating heart rate. *J Intensive Care*. (2020) 8:41.
7. Karim HMR, Esquinas AM. Success or failure of high-flow nasal oxygen therapy: the ROX Index is good, but a modified ROX index may be better. *Am J Respir Crit Care Med*. (2019) 200(1):116–7.
8. Yarnell CJ, Johnson A, Dam T, et al. Do thresholds for invasive ventilation in hypoxemic respiratory failure exist? A cohort study. *Am J Respir Crit Care Med*. (2023) 207(3):271–82.
9. Yu H, Saffaran S, Tonelli R, et al. Machine learning models compared with current clinical indices to predict the outcome of high-flow nasal cannula therapy in acute hypoxemic respiratory failure. *Crit Care*. (2025) 29(1):101.
10. Badawy J, Nguyen OK, Clark C, et al. Is everyone really breathing 20 times a minute? Assessing epidemiology and variation in recorded

respiratory rate in hospitalized adults. *BMJ Qual Saf*. (2017) 26(10):832–6.
11. Brakenhoff TB, Mitroiu M, Keogh RH, et al. Measurement error is often neglected in medical literature: a systematic review. *J Clin Epidemiol*. (2018) 98:89–97.
12. Philip KE, Pack E, Cambiano V, Rollmann H, Weil S, O'Beirne J. The accuracy of respiratory rate assessment by doctors in a London teaching hospital: a cross-sectional study. J Clin Monit Comput 2015; 29: 455–60.
13. Lauteslager T, Dishakjian V, Watson L, et al. Detecting early signs of deterioration and preventing hospitalizations in skilled nursing facilities using remote respiratory monitoring. *Respir Med Case Rep*. (2024) 50:102044.
14. Drummond GB, Fischer D, Arvind DK. Current clinical methods of measurement of respiratory rate give imprecise values. *ERJ Open Res*. (2020) 6(3):00023-2020.
15. Smith SF, Duell DJ, Martin BC. Clinical Nursing Skills. 8th ed. Upper Saddle River, NJ: Pearson; 2012.
16. Hollerbach AD, Sneed NV. Accuracy of radial pulse assessment by length of counting interval. Heart Lung. 1990 May;19(3):258-64. PMID: 2341264.
17. Bent, B., Goldstein, B.A., Kibbe, W.A. *et al.* Investigating sources of inaccuracy in wearable optical heart rate sensors. *npj Digit. Med.* **3**, 18 (2020). https://doi.org/10.1038/s41746-020-0226-6
18. Hollmann, N., Müller, S., Purucker, L. et al. Accurate predictions on small data with a tabular foundation model. Nature 637, 319–326 (2025). https://doi.org/10.1038/s41586-024-08328-6
19. Brown, T. et al. Language models are few-shot learners. In Proc. Advances in Neural Information Processing Systems (eds Larochelle, H. et al.) Vol. 33, 1877–1901 (Curran Associates, 2020).
20. Yu, H., Saffaran, S., Maia, I.S. *et al.* Early prediction of non-invasive ventilation outcome using the TabPFN machine learning model: a multi-centre validation study. *Intensive Care Med* (2025). https://doi.org/10.1007/s00134-025-08025-6
21. RENOVATE Investigators and the BRICNet Authors. High-flow nasal oxygen vs noninvasive ventilation in patients with acute respiratory failure: The RENOVATE randomized clinical trial. JAMA. 2024.
22. Tonelli R, Fantini R, Bruzzi G, et al. Effect of high flow nasal oxygen on inspiratory effort of patients with acute hypoxic respiratory failure and do not intubate orders. Intern Emerg Med. 2024;19:333–42.
23. Tonelli R, Cortegiani A, Fantini R, et al. Accuracy of nasal pressure swing to predict failure of high-flow nasal oxygen in patients with acute hypoxemic respiratory failure. Am J Respir Crit Care Med. 2023;207(6):787–9.
24. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely

accessible electronic health record dataset. Sci Data. 2023.
25. Pollard T, Johnson A, Raffa J, et al. The eICU Collaborative research database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178.
26. Noble W. What is a support vector machine?. Nat Biotechnol. 2006; 24, 1565–1567.
27. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015; 27(2):130-5.
28. Ontivero-Ortega M, Lage-Castellanos A, Valente G, Goebel R, Valdes-Sosa M. Fast Gaussian Naïve Bayes for searchlight classification analysis. Neuroimage. 2017; 163:471-479.
29. Friedman J H. Stochastic gradient boosting. Comput Stat Data Anal. 2002, 38(4): 367-378.
30. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc KDD ACM. 2016; 2016: 785–794.
31. Liaw R, Liang E, Nishihara R, et al (2018) Tune: a research platform for distributed model selection and training. arXiv.
32. Lundberg SM, Erion G, Chen H et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020; 2: 2522–5839.
33. Poncet A, Perneger TV, Merlani P, Capuzzo M, Combescure C. Determinants of the calibration of SAPS II and SAPS 3 mortality scores in intensive care: a European multicenter study. Crit Care. 2017;21:1-10.
34. Kobayashi H. Effect of measurement duration on accuracy of pulse-counting. Ergonomics. 2013;56(12):1940-4. doi: 10.1080/00140139.2013.840743. Epub 2013 Oct 11. PMID: 24117167; PMCID: PMC3877911.
35. Karaivanova A, Ivanovska S, Gurov T. Monte Carlo method for density reconstruction based on insufficient data. Procedia Comput Sci. (2015) 51:1782–90.
36. Drummond GB, Bates A, Mann J, Arvind DK. Characterization of breathing patterns during patient-controlled opioid analgesia. Br J Anaesth. (2013) 111(6):971–8
37. Lee PJ. Clinical evaluation of a novel respiratory rate monitor. J Clin Monit Comput. (2016) 30(2):175–83. doi: 10.1007/s10877-015-9697-4
38. PMD Solutions. RespiraSense: Continuous Respiratory Rate Monitoring. PMD Solutions. (2025) Available from: https://www.pmd-solutions.com/product/
39. Subbe CP, Kinsella S. Continuous Monitoring of Respiratory Rate in Emergency Admissions: Evaluation of the RespiraSense™ Sensor in Acute Care Compared to the Industry Standard and Gold Standard. Sensors (Basel). (2018) 17;18(8):2700.

40. Drummond GB, Fischer D, Lees M, Bates A, Mann J, Arvind DK. Classifying signals from a wearable accelerometer device to measure respiratory rate. ERJ Open Res. (2021) 7(2):00681-2020.
41. Drummond GB, Bates A, Mann J, et al. Validation of a new non-invasive automatic monitor of respiratory rate for postoperative subjects. Br J Anaesth. (2011) 107(3):462– 9.
42. Miechels J, Koning MV. Respiratory rate measurement by pressure variation in the high flow nasal cannula-system in healthy volunteers. (2024) J Clin Monit Comput 38, 1397–1404.
43. Bent, B., Goldstein, B.A., Kibbe, W.A. et al. Investigating sources of inaccuracy in wearable optical heart rate sensors. npj Digit. Med. 3, 18 (2020). https://doi.org/10.1038/s41746-020-0226-6
44. van der Ploeg, T., Austin, P.C. & Steyerberg, E.W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* **14**, 137 (2014). https://doi.org/10.1186/1471-2288-14-137
45. Gorishniy, Y., Kotelnikov, A., & Babenko, A. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. (2024) *arXiv preprint arXiv:2410.24210*.
46. Pavlou, M., Ambler, G., Qu, C. *et al.* An evaluation of sample size requirements for developing risk prediction models with binary outcomes. *BMC Med Res Methodol* **24**, 146 (2024). https://doi.org/10.1186/s12874-024-02268-5