Non-invasive Total Serum Bilirubin Estimation in Preterm Infants with Modified Mixed-Effects Random Forest

Meng Chen[1], Alain Beuchée[1], Fabrice Tudoret[1], Alfredo I Hernández[1]

[1] University of Rennes, CHU Rennes, Inserm, LTSI-UMR 1099, Rennes, France

Abstract

This study presents the development of non-invasive total serum bilirubin (TSB) estimators using modified mixed-effects random forest (MERF) models, which incorporate heart rate variability (HRV) and physiological insights. We analyzed data from 326 preterm infants, comprising 1,652 TSB samples. A set of HRV parameters centered on TSB measurements was obtained through a processing pipeline including data querying, ECG signal processing, RR series estimation and feature engineering. Results demonstrated that (m)MERF models outperform standard random forests (RF) by achieving higher Pearson correlations, higher agreements and lower proportional bias. While the models require patient-specific historical data for initialization, the proposal of non-invasive TSB estimators shows clinical potential to minimize blood exploitations, enhance real-time TSB destination and optimize interventions for hyperbilirubinemia, offering a promising step towards personalized and less invasive neonatal care.

1. Introduction

Hyperbilirubinemia affects 60% to 80% of newborns [1] and poses significant risks to preterm infants. Prematurity and underdeveloped hepatic function lead to bilirubin accumulation, causing irreversible neurotoxicity and conditions such as kernicterus if untreated [2]. Preterm infants with higher peak bilirubin values after birth are at greater risk of bilirubin-induced neurotoxicity and neurodevelopmental impairment compared to term infants [3–5]. Regular total serum bilirubin (TSB) monitoring is crucial for timely interventions including phototherapy and exchange transfusion. However, it typically involves repetitive blood sampling, which is invasive and painful for those tiniest babies and may lead to complications such as anemia, infection and osteomyelitis [6]. Non-invasive alternatives in order to optimize blood sampling sensing instants and ultimately minimize blood exploitation and associated risks are now receiving much attention.

Research has shown that hyperbilirubinemia affects the autonomic nervous system (ANS) activities, especially parasympathetic predominance, manifested by alterations in heart rate variability (HRV) in jaundiced term and preterm neonates [7–9]. Since these preterm newborns are continuously monitored, including continuous acquisitions of ECG signals, we hypothesize that the joint analysis of HRV, from the monitoring ECG, and basic clinical information could be used to estimate bilirubin levels using signal processing and machine learning (ML) models.

The selection of the appropriate ML algorithm is a challenge in this field, characterized by longitudinal, time-dependent and noisy data. Longitudinal data exhibits temporal dependencies between observations, making it difficult for traditional ML models that assume independence. Covariates may change over time, requiring their inclusion in the model. The non-stationarity of the data can also impact model performance. Furthermore, processing longitudinal data may result in low model interpretability, if the selected model does not allow for the understanding of how it uses information from different time points and how this affects predictions.

To address these issues, we proposed novel TSB estimators based on mixed-effects random forests, incorporating specific physiological insights as mixed-effects terms. We evaluated and compared these models with a standard random forest.

2. Methods

2.1. Clinical study and population

We conducted an ancillary study of a prospective multicenter clinical study CARESS-Premi (NCT01611740) for the evaluation of estimating TSB levels based on known longitudinal cardio-respiratory data and general clinical information. The CARESS-Premi cohort involved preterm infants born between 24 and 32 weeks gestational age (GA), with birth weights >500 grams, postnatal age (PNA)>3 days, post-menstrual age (PMA)<34 weeks and admitted to the NICU across three university hospitals in France (University Hospitals of Rennes, Angers and Lille) between October 2012 and November 2018. The protocol was conducted with approval by the local ethics committee

and informed parental consent was obtained.

All infants were continuously monitored for their cardio-respiratory status (IntelliVue MP40 Philips Medical System, Eindhoven, Netherlands). Cardio-respiratory signals were acquired from monitoring systems and de-identified and transferred to an original cloud-based system, ASCENT, for further analysis. A subset of the whole cohort with bilirubin-related conditions during follow-ups is eligible for this study. Patients with TSB levels above 400 μmol/L were excluded as these extremely high levels often due to complicated conditions associated with both immaturities and pathologies that may not accurately reflect the typical clinical scenarios we aim to study. Blood samples used to determine TSB levels were collected for routine clinical care and adhered to standard care criteria.

2.2. Data processing and feature extraction

1) Data selection: Bilirubin levels were sparsely measured from each patient. To acquire valuable data, we first queried the instants of TSB measurements. We then selected four hours of ECG recordings and associated clinical data centered on each TSB sampling (two hours before and two hours after). Further processing was performed on these 4-hour data segments.

2) ECG signal processing and data denoising: Raw ECG signals were segmented into successive 15-minute windows with 20% overlap. Each segment was pre-processed to detect cardiac beats (QRS complex) using a robust real-time QRS detector adapted to specific characteristics of the newborn electrophysiology, previously proposed and evaluated by our team [10]. RR intervals were calculated from detected cardiac cycles. A multi-step approach using logic rules based on pathology and rhythm correction was applied to automatically reject or correct artifacts and errors. Changes over time in the mean and variance of the corrected series were estimated, and signal stationarity was analyzed for each 15-minute segment. The most stationary segment of the 4-hour corrected RR series was used for further analysis.

3) Feature extraction: HRV features describing cardiovascular functions modulated by the ANS were extracted from the most stationary RR series. This included 12 time-domain features: mean, median, SD, RMSSD, skewness, kurtosis, deceleration capacity and acceleration capacity, inter-decile range between 10^{th} and 90^{th} percentile, percentage deceleration of RR intervals (pDec), standard deviation of RR corresponding to pDec and sample asymmetry. The five frequency-domain features are: power in the low-frequency spectral band (LF, 0.02-0.2 Hz), normalized LF, high-frequency power (HF, 0.2-2 Hz), normalized HF, and LF/HF ratio. Finally, 5 nonlinear features include SD1 and SD2 from the Poincaré plot, sample entropy, and detrended fluctuation analysis ($\alpha 1$ and $\alpha 2$). Outliers in the HRV feature space were then detected and excluded using a Chi-square test (α=0.05) on Mahalanobis distance.

2.3. Machine Learning models for clinical longitudinal data analytics

Machine learning models are widely used in the field of healthcare. However, classical ML algorithms face challenges when dealing with longitudinal data, which consists of repeated clinical information from the same patients over an extended period. Starting with a typical model of random forest, we compared several variants of mixed-effects-based ML models, which are well-suited to this genre of problem, particularly when the random effects are non-negligible [11]. **1) Baseline random forest (RF):** Standard RF regressors were employed for TSB estimation in the form of

$$y = f(\boldsymbol{X}) + \epsilon \tag{1}$$

where y is the TSB levels to estimate, $\boldsymbol{X}$ the features comprising 22 HRV parameters and 2 age indicators (PMA and PNA), and ϵ the estimation error.

2) Mixed-effects random forest (MERF): The MERF, originated from linear mixed-effects models, has been extended to better handle clustered and longitudinal data with repeated measurements within clusters/subjects. Hajjem et al. [11] proposed in 2014 the MERF integrating a powerful ensemble learning method, formulated as

$$y = f(\boldsymbol{X}) + \boldsymbol{b_i Z} + \boldsymbol{\epsilon} \tag{2}$$

where y is TSB measurements, $\boldsymbol{X}$ is a matrix of fixed-effects features, $f(\cdot)$ denotes a nonlinear function, i.e., a standard RF regressor, $\boldsymbol{Z}$ is a matrix of random-effects covariates specific to each subject capturing inter-subject variability, $\boldsymbol{b_i} \sim N(0, \boldsymbol{D})$ are random effects coefficients for each patient i, and $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{R_i})$ represents individual error. The covariance matrix of $\boldsymbol{b_i}$ is $\boldsymbol{D}$, while $\boldsymbol{R_i}$ is the covariance matrix of $\boldsymbol{\epsilon}$.

The MERF is implemented with the expectation-maximization (EM) algorithm, which was originally used for linear mixed-effects models. This algorithm iterates until convergence, jointly learning random effects coefficients, error priors and the RF, evaluating the generalized log-likelihood (GLL) criterion at each iteration.

3) Modified mixed-effects random forest (mMERF): In our previous work [12], we observed that the natural evolution of TSB levels with PNA follows an exponential decay pattern. To capture this nonlinear relationship explicitly, we introduce a modified MERF model:

$$y = f(\boldsymbol{X}) + b1_i(Z1_i) + b2_i(g(Z2_i)) + \epsilon \tag{3}$$

where y represents TSB measurements, $f(X)$ is a standard RF model, $b1_i$ and $b2_i$ are coefficients for two random effects specific to patient i: $Z1$ (PMA) and $Z2$ (PNA), and $g(\cdot)$ is an exponential decay function characterizing the overall TSB dynamics:

$$g(\boldsymbol{Z2_i(t)}) = A \times exp(-B \cdot \boldsymbol{Z2_i(t)} + GA + Tc) + C \tag{4}$$

where $\boldsymbol{Z2_i(t)}$ are the PNA in days when bilirubin levels were measured, coefficients A (118.34), B (0.1142), C (0), Tc (-201.96), GA (196.00) are the median values obtained from 72 patient-specific TSB decay models recently proposed in [12].

4) Model development: We developed models using a feature set of extracted HRV and clinical indicators, with observed TSB levels as targets. Data were shuffled and split into a training set (80%) for learning and a test set (20%) for validating without differentiating subjects.

Four sets of variants were developed to evaluate the added value of the proposed random-effects factors:
(i) a standard RF used as a baseline model (RF_{base});
(ii) pure MERF with no extra random effects but only assigning clusters (infants) for samples ($MERF_0$);
(iii) MERF with two random effects of PMA and PNA to capture patient-level variability ($MERF_2$);
(iv) a modified MERF with a linear random effect of PMA and a nonlinear random effect of PNA (mMERF). The RFs integrated into the (m)MERFs are marked as $RF_{(m)MERF}$.

Optimal hyper-parameters for the baseline RF and the RFs embedded in (m)MERFs were identified via Bayesian optimization within predefined parameter spaces. The maximum iteration of the EM for learning (m)MERF models was set to 200. We utilized an open-source Python package of *MERF* [11] for the modeling process.

5) Evaluation metrics: The fitted (m)MERF estimators work as follows: for new observations from patients seen during training ("known" patients), predictions comprise population-averaged RF estimations with random effects corrections; for samples from patients not seen during training ("unknown" patients), the estimators revert to standard RFs with only fixed effects.

Pearson correlation and the root-mean-squared error (RMSE) were used to quantify the correlation between the observed TSB levels and the estimated levels. Bland-Altman (B&A) analysis and plots were used to compare the agreements between observed and estimated TSB measurements, and 95% limits of agreements (LoA) were calculated. In addition, based on the B&A plots, we calculated the correlation between the differences and the means of differences and applied linear regression to fit a bias line to quantify the consistency and discrepancies.

3. Results

A total of 326 infants and 1,652 TSB samples were initially included in the study. After processing and exclusions, 1,477 samples from 319 patients remained for analysis. The median number of observations per patient was 4 (ranging from 1 to 22), and the mean TSB level was 118±50 μmol/L.

Table 1 presents the models' overall performances. In general, from RF_{base} to $MERF_0$, $MERF_2$ and finally mMERF, the model performances improved as the model structure expanded and incorporated information increased. In terms of the Pearson correlations and RMSE between the observed and estimated TSB levels, the proposed three models achieved higher correlations (above 0.80) and lower estimation error (around 30 μmol/L). Regarding the B&A analysis, the mixed-effects-based models enhanced the agreements by narrowing LoA by approximately 30% on the test data (around 50% on the training data). Besides, Pearson correlations and linear regressions were computed between the means and differences of observed and model-estimated TSB levels to quantify the estimators' proportional bias. The results show that models of (m)MERFs greatly attenuated the undesirable bias by decreasing the linear regression slopes from 0.76 to 0.24 for the samples from "known" patients on the test set (from 0.53 to 0.13 on the training set). In addition, it is noticed that the RFs embedded in the (m)MERFs, marked as $RF_{(m)MERF}$, have even poorer performance relative to the RF_{base}, which is reasonable.

Table 1: Model performances on the test set.

	Corr.†	RMSE (μmol/L)	LoA	ΔLoA	Corr. diff‡	Linear bias ⋆
RF_{base}	0.53	42.41	83.10	/	0.61	0.76
RF_{MERF_0}	0.51	43.64	84.51	+1.70%	0.63	0.81
$MERF_0$	0.80	30.12	58.88	-29.15%	0.41	0.29
RF_{MERF_2}	0.50	44.00	84.92	+2.19%	0.64	0.83
$MERF_2$	0.83	28.40	55.58	-33.12%	0.42	0.27
RF_{mMERF}	0.49	44.62	85.87	+3.33%	0.59	0.77
mMERF	0.83	28.20	55.21	-33.56%	0.39	0.24

ΔLoA: Changes in LoA relative to the RF_{base} model.
†Pearson correlation between the observed and estimated TSB levels.
‡Pearson correlation between the mean and differences of observed and estimated TSB levels of samples from "known" patients.
⋆The slope in linear regression of samples from "known" patients.

Fig. 1 illustrates the performance comparison of RF_{base} and mMERF model on the training and test set. RF_{base} exhibited wider dispersion and evident proportional bias on both train and test data, while the mMERF effectively improved the estimation.

4. Discussion

This study concerns the assessment of the added value of novel mixed-effects-based models in estimating TSB levels in preterm infants when incorporating physiological insights. By combining different random effects covariates, the results show that the proposed MERF and mMERF models outperform the standard RF baseline by achieving higher correlations, narrower LoAs and lower proportional bias, suggesting their advantages in analyzing longitudinal clinical data with patient-level repeated measurements. Furthermore, integrating a knowledge-based function describing TSB evolution in preterm infants as additional random effect is useful in optimizing the estimators.

However, this work has some limitations. While the

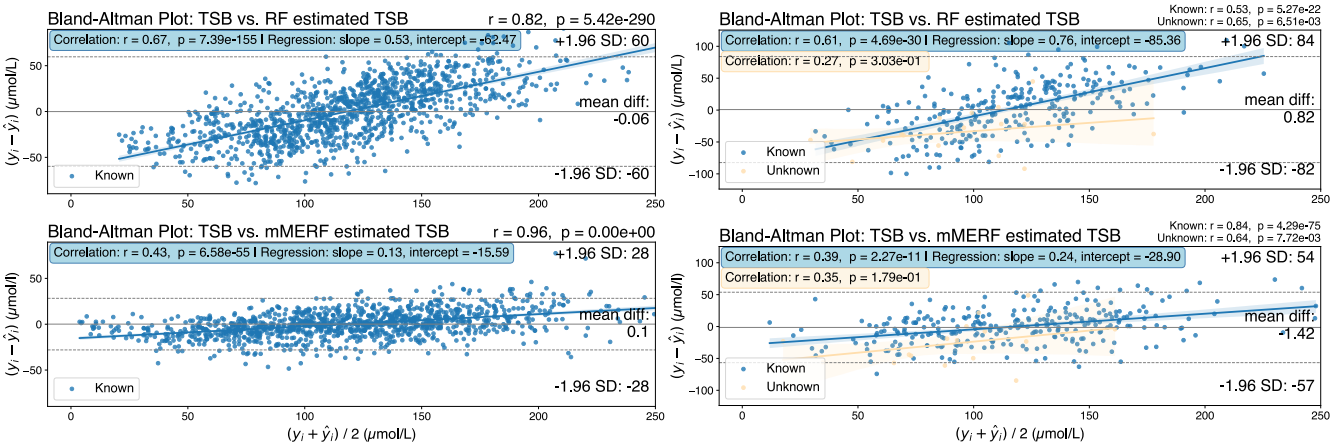

Figure 1: Bland-Altman plots comparing RF_{base} (top) and mMERF (bottom) models concerning the differences between observed (y_i) and estimated ($\hat{y_i}$) TSB levels. Left: on the training set. Right: on the test set.

(m)MERF models showed greater consistency and less bias in TSB estimation for known" patients, there is no significant difference in performance over the baseline models when encountering unknown" patients in the test set. This is due to the inherent nature of mixed-effects models, which require patient-specific historical data to initialize the random coefficients. In addition, the estimations from the proposed models still presented some deviations from observed distributions, possibly due to the limited sample size of a subset of the clinical proposal. Further improvements could involve more systematic and extensive data collection to generate larger datasets and broader validation. Incorporating new physiological parameters might also enhance the model's performance.

5. Conclusion

An original set of models incorporating personalized and knowledge-based random effects was proposed for non-invasive TSB estimation in preterm infants. The mixed effects MERF models improved agreements and reduce the proportional bias by explicitly integrating meaningful physiological insights. Though they require patient-specific historical data for initialization, with a larger dataset and broader validation, these models could be integrated into NICU monitoring systems to enhance real-time TSB estimation and optimize hyperbilirubinemia interventions, offering a promising step towards personalized and less invasive neonatal care.

Acknowledgments

This study was partly funded by the China Scholarship Council.

References

[1] NICE. Jaundice in newborn babies under 28 days | Guidance. https://www.nice.org.uk/guidance/cg98, 2010. Accessed: 2024-06-26.

[2] Johnson L, Bhutani VK. The clinical syndrome of bilirubin-induced neurologic dysfunction. In Seminars in perinatology, volume 35. Elsevier, 2011; 101–113.

[3] Gartner LM, Snyder RN, Chabon RS, Bernstein J. Kernicterus:high incidence in premature infants with low serum bilirubin concentrations. Pediatrics 1970;45(6):906–917.

[4] Wusthoff CJ, Loe IM. Impact of bilirubin-induced neurologic dysfunction on neurodevelopmental outcomes. In Seminars in Fetal and Neonatal Medicine, volume 20. Elsevier, 2015; 52–57.

[5] Solis-Garcia G, Raghuram K, Augustine S, Ricci MF, St-Hilaire M, Louis D, Makary H, Yang J, Shah PS. Hyperbilirubinemia among infants born preterm: Peak levels and association with neurodevelopmental outcomes. The Journal of Pediatrics 2023;259:113458.

[6] El-Beshbishi SN, Shattuck KE, Mohammad AA, Petersen JR. Hyperbilirubinemia and transcutaneous bilirubinometry. Clinical chemistry 2009;55(7):1280–1287.

[7] Uhrikova Z, Zibolen M, Javorka K, Chladekova L, Javorka M. Hyperbilirubinemia and phototherapy in newborns: Effects on cardiac autonomic control. Early Human Development 2015;91(6):351–356.

[8] Specq ML, Bourgoin-Heck M, Samson N, Corbin F, Gestreau C, Richer M, Kadhim H, Praud JP. Moderate hyperbilirubinemia alters neonatal cardiorespiratory control and induces inflammation in the nucleus tractus solitarius. Frontiers in physiology 2016;7:437.

[9] Al-Omar S, Le Rolle V, Samson N, Specq ML, Bourgoin-Heck M, Costet N, Carrault G, Praud JP. Influence of moderate hyperbilirubinemia on cardiorespiratory control in preterm lambs. Frontiers in physiology 2019;10:468.

[10] Doyen M, Ge D, Beuchée A, Carrault G, I. Hernández A. Robust, real-time generic detector based on a multi-feature probabilistic method. PLoS ONE 2019;14(10):e0223785.

[11] Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 2014;84(6):1313–1328.

[12] Chen M, Beuchée A, Levine E, Storme L, Gascoin G, Hernández AI. Model-based characterization of total serum bilirubin dynamics in preterm infants. Pediatr Res ;In Press.

Address for correspondence:

Alfredo I. Hernández

Univ Rennes, Inserm, LTSI-UMR 1099, F-35000 Rennes, France

alfredo.hernandez@univ-rennes.fr