Massive data processing and explainable machine learning in neonatal intensive care units

Meng Chen

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Signal, Image, Vision*

Par
Meng CHEN

Massive Data Processing and Explainable Machine Learning in Neonatal Intensive Care Units

Thèse présentée et soutenue à Rennes, le 18 décembre 2024
Unité de recherche : Laboratoire Traitement du Signal et de l'Image (LTSI) - UMR INSERM 1099

Rapporteurs avant soutenance :

Olivier MESTE Professeur des Universités, Université Côte d'Azur, France
Julien OSTER Chargé de Recherche INSERM, Université de Lorraine, France

Composition du Jury :

Président :	Olivier MESTE	Professeur des Universités, Université Côte d'Azur, France
Examinateurs :	Olivier MESTE	Professeur des Universités, Université Côte d'Azur, France
	Julien OSTER	Chargé de Recherche INSERM, Université de Lorraine, France
	Julie FONTECAVE-JALLON	Maître de Conférences, Université Grenoble Alpes, France
	Alain BEUCHÉE	Professeur des Universités et Praticien Hospitalier, Univeristé de Rennes
Dir. de thèse :	Alfredo I. HERNÁNDEZ	Directeur de Recherche INSERM, Univeristé de Rennes
Co-dir. de thèse :	Lotfi SENHADJI	Professeur des Universités, Univeristé de Rennes

献给我的父母、胞弟与爱人

Acknowledgement

I would like to express my appreciation to the members of my jury for accepting the invitation to endorse my Ph.D. and for traveling from afar to be physically present at my Ph.D. defense. It was such a great honor for me. I appreciate Prof. Olivier Meste and Dr. Julien Oster for their high-quality reports with their inspiring comments, insightful questions and constructive suggestions. I appreciate equality to Associate Prof. Julie Fontecave-Jallon for serving as both an examiner and a CSI member during my PhD years and providing valuable comments and help.

I would like to deeply thank my Ph.D. supervisors, Alfredo Hernández and Lotfi Senhadji, as well as Alain Beuchée, whose expertise and guidance made these three years of research inspiring and rewarding. Their unwavering commitment to my projects and their respective skills were the main drivers of my achievements. I am particularly grateful to Alfredo for his invaluable advice, his constant kindness and for always being present at the right moments. From my short internship in 2019, I already realized how exceptional he is : as a researcher, a mentor, a leader and simply as someone whose experience and wisdom I deeply admire. His recognition of my potential and offer to undertake this Ph.D. journey was where it all began. During my doctoral studies, we maintained very good communication and supervision at just the right pace (neither oppressive nor laissez-faire). I always learned new perspectives from him during our exchanges. He has been both my mentor and my steadfast support throughout this journey, and for that, I am forever grateful. Lotfi is indispensable for my stay in LTSI in a more general sense. His hospitality and patience as well as his deep connection to China always made me feel at home; likewise, I was very grateful for his presence and help during my defense preparation. As for Alain, as an excellent front-liner in pediatrics, his involvement with expertise and encouragement brought nuance to my research, bridging it with clinical practice. He surrounded me with a remarkable clinical team, fostering collaboration and providing invaluable insights, which greatly enriched my work.

It has been an immense privilege to work with the passionate and inspiring SEPIA team. My thesis stands on the shoulders of my predecessors and their significant contributions. I would like to thank Oscar Acosta Tamayo for his thoughtful comments and affirmation of my work as a CSI committee member. I also appreciate Fabrice Tudoret for his technical support, collaboration, and exceptional sense of humor.

I extend my gratitude to the successive directors of LTSI, Lotfi Senhadji, Fabrice Wendling, and Mireille Garreau, for fostering a welcoming and stimulating environment in the laboratory. My

heartfelt thanks also go to the administrative team, including Patricia Bernabé, Soizic Charpentier, Muriel Diop, and Servane Le Guyader, whose professionalism and kindness often warmed my heart.

A big thank you to all the members of the laboratory who made coming to work a pleasure. I am particularly grateful for the camaraderie of such a great team of doctoral students, postdoctoral fellows, and interns, with whom I shared both work and laughter-filled lunch breaks. I would like to thank my officemates Orlane Duport, Marion Taconné, Salman Almuhammad Alali, and Francesca Menna for the pleasant working time and many rich exchanges. A special shoutout to the Durak Crew, no one would ever predict that a group of us would go so far. *Benninging* with playing cards, this cute crew gradually grew into a supportive and warm family where we spent so much quality time and shared a lot of happiness, encouragement, and, most importantly, love. Thank you so so much for gathering together and being friends, and I wish our friendship lasts longer and longer. Especially, Houda Jebbari, I am so lucky to have shared many beautiful and unforgettable memories with you in many places in the world that I will always cherish. During these moments, we exchanged our thoughts and cared for each other, and our similarities with just the right amount of differences make us friends. You are such a wonderful girl with a beautiful personality, thank you for your existence and for everything.

I would like to thank my Chinese colleagues and friends at LTSI, Fuzhi Wu, Chen Zhang, Yang Li, Zuyi Yu, Qixiang Ma, Shun Lyu and many others, their help, companionship and encouragement were a source of comfort. They brought the warmth of home to a foreign land. I especially thank Chen and Fuzhi for being close friends and creating an amazing utopia where we could share and express our feelings or topics, huge or tiny. I also thank Qixiang for being a big brother to us.

I would thank my loved ones, my family (Mama, Baba, and little Didi), my partner Jia, thank you so much for your unwavering encouragement and support every step of this adventure, and thank you for your unconditional love. Jia, distance and time have never been an obstacle between us, your constant support over these past seven years has been my anchor. You are my best friend, soulmate, and greatest love. I am deeply grateful for the journey we have shared and look forward to an even brighter future together.

Words will never be enough to express my gratitude to everyone who supported me throughout this journey. I may have missed some names in this short acknowledgment, but all the people I have met and all the experiences I had in France during these 39 months have shaped me in profound ways. I think this is what we called growth. All in all, I am infinitely grateful for all, all those who contributed directly or indirectly to my journey, and I sincerely wish you happiness and success in all that lies ahead.

Résumé en français

Les nouveau-nés prématurés sont définis comme des enfants nés avant 37 semaines complètes de gestation [1]. Dans le monde, environ 15 millions de nouveau-nés sont nés avant terme, ce qui équivaut à 1 nouveau-né sur 10 [2]. En France, on dénombre 122 naissances prématurées par jour et 44 500 naissances prématurées sur l'année, ce qui représente 7,0 % des naissances [3]. La naissance précoce interrompant le développement complet de leurs systèmes physiologiques, ces nouveau-nés prématurés sont particulièrement fragiles et sujets à diverses complications, telles que des affections cardio-respiratoires, immunologiques, neurologiques et digestives. Les complications liées à la prématurité restent le principal facteur de mortalité des enfants de moins de 5 ans [2, 4] et sont responsables de plus de la moitié de la morbidité à long terme [5].

Les soins néonatals, en particulier dans les unités de soins intensifs néonatals (USIN), sont essentiels pour la survie et le développement des nouveau-nés prématurés et gravement malades. Face à leur vulnérabilité, les USIN offrent à ces nouveau-nés fragiles des soins médicaux spécialisés et continus, grâce à des systèmes spécialisés de maintien en vie, des technologies de surveillance avancées et des interventions cliniques adaptées. Cependant, malgré les progrès en médecine néonatale et dans les pratiques de soins, la gestion des conditions à haut risque dans les USIN reste confrontée à des défis importants qui sont systématiquement associés à des séjours prolongés à l'hôpital et qui ont des répercussions sur les résultats des patients.

L'amélioration des résultats de santé des nourrissons prématurés repose sur la capacité à diagnostiquer et intervenir sur des conditions critiques le plus tôt possible. La détection précoce d'événements indésirables, notamment par des méthodes non invasives, reste un défi mais est devenue une tendance essentielle dans les soins néonatals. Au cours des dernières décennies, une variété de systèmes d'aide à la décision clinique (CDSS), qui exploitent les données massives de santé, les informations fondées sur des preuves et l'analyse statistique ou les techniques d'intelligence artificielle, ont été développés pour augmenter les capacités de prise de décision des professionnels médicaux dans les soins néonatals [6–9]. Notre équipe de recherche est particulièrement visible sur ces activités [10–18].

Cependant, l'intégration de tels systèmes dans les flux de travail cliniques pose des défis importants, notamment en ce qui concerne la complexité de l'acquisition et du traitement des données, le besoin d'explicabilité des modèles et la difficulté de mise en uvre dans le monde réel [19, 20]. Relever ces défis est crucial pour le succès du déploiement des CDSS dans les USIN, où

des quantités massives de données de surveillance peuvent être transformées en informations exploitables, améliorant ainsi les résultats à court et à long terme pour les nouveau-nés prématurés.

En particulier, cette thèse s'intéresse à un ensemble de **défis méthodologiques majeurs** inhérents au travail avec des modèles d'apprentissage automatique sur les données de surveillance néonatale en USIN, qui sont longitudinales, continues, dépendantes du temps et souvent bruitées. Les principaux défis souvent négligés dans ce contexte incluent :

- Dépendances temporelles et covariables évolutives : Les données longitudinales présentent souvent des dépendances temporelles entre les observations, et les covariables (variables prédictives) peuvent évoluer au fil du temps, ce qui limite l'application des méthodes classiques d'ingénierie des caractéristiques et d'apprentissage automatique, qui supposent généralement l'indépendance.
- Données manquantes : Les données longitudinales contiennent souvent des valeurs manquantes en raison d'erreurs de mesure, de transmission ou du bruit, ce qui peut entraîner des estimations biaisées si elles ne sont pas correctement traitées.
- Granularité temporelle et non-stationnarité : Les données peuvent être collectées à différentes échelles temporelles et peuvent présenter des fortes périodes de non-stationnarité.
- Complexité du modèle et évolutivité : Le caractère longitudinal des données peut accroître la complexité du modèle, augmentant ainsi le risque de surapprentissage et les coûts computationnels, ce qui exige des techniques efficaces de régularisation et la proposition de modèles évolutifs.
- Interprétabilité des modèles : Il est inmportant de comprendre comment les modèles utilisent les informations temporelles et comment cela affecte les prédictions, afin de s'assurer que celles-ci soient significatives et exploitables.

En réponse à ces défis, l'objectif de cette thèse de doctorat est de proposer des nouvelles méthodes de traitement et des nouveaux outils applicables au processus de prise de décision impliqué dans l'unité de soins intensifs néonatals, d'une manière non invasive, continue et en temps pseudo-réel. Plus particulièrement, ces travaux visent à :

- Améliorer le diagnostic et l'optimisation des thérapies pour les nourrissons prématurés en analysant des caractéristiques dynamiques telles que la variabilité de la fréquence cardiaque (HRV), extraites de signaux physiologiques (par exemple, des enregistrements cardio-respiratoires) à l'aide de méthodes spécifiques et avancées de traitement des données et d'algorithmes d'apprentissage automatique.
- Approfondir la compréhension des schémas physiologiques chez les nourrissons prématurés en analysant des données longitudinales et en les corrélant avec les résultats cliniques à travers des approches basées sur des modèles et des connaissances.
- Intégrer les modèles et algorithmes proposés dans un système d'aide à la décision clinique, permettant des inférences en quasi-temps réel pour assister les cliniciens en fournissant des informations opportunes et fondées sur les données.

Cette thèse a été réalisée dans le cadre d'une étude clinique prospective nationale multicentrique CARESS-Premi (NCT01611740), "Contribution des analyses en temps réel des signaux CARdio-RESpiratoires au diagnostic d'infection chez les prématurés", et elle concerne environ 500 prématurés recrutés entre octobre 2012 et novembre 2018 à travers les USIN de trois centres hospitaliers universitaires français (Rennes, Lille et Angers). Dans cette thèse, deux défis cliniques les plus fréquentes et les plus difficiles en USIN ont été particulièrement étudiés : l'hyperbilirubinémie néonatale (jaunisse) [21, 22] et le sepsis tardif néonatal (infection nosocomiale) [23–26].

La variabilité de la fréquence cardiaque (HRV), dérivée des signaux cardiaques, constitue un outil clé pour évaluer la régulation du système nerveux autonome chez les nourrissons prématurés dans diverses conditions physiologiques et pathologiques. Le diagnostic reposant sur l'ECG et la HRV présente l'avantage d'être non-invasif et facilement disponible en continu dans le contexte des unités de soins intensifs néonatals (USIN). Il offre des informations précieuses sur l'impact de l'hyperbilirubinémie et du sepsis chez les nouveau-nés à un niveau plus approfondi. Des niveaux élevés de bilirubine ont été rapportés comme ayant une influence sur la fonction autonome chez les nouveau-nés à terme et prématurés atteints de jaunisse, modifiant les schémas de fréquence cardiaque. Certains paramètres spécifiques de la variabilité de la fréquence cardiaque peuvent révéler ces effets, offrant ainsi des marqueurs prédictifs potentiels pour l'hyperbilirubinémie [27–30]. De même, le sepsis induit une inflammation systémique qui se manifeste souvent par une régulation autonome anormale. Des études antérieures ont montré que l'analyse de la HRV peut servir d'indicateur précoce de la progression du sepsis, fournissant des informations sur la réponse de l'organisme à l'infection avant même l'apparition des symptômes cliniques [31–35].

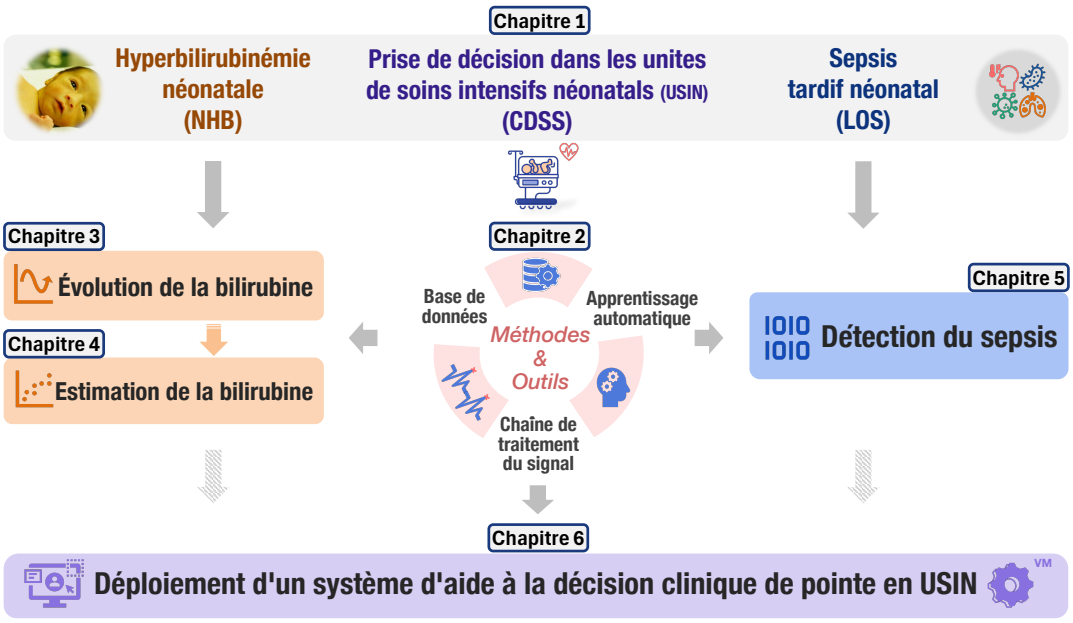

FIGURE 1 : Cadre d'étude de la thèse.

Dans ce contexte, les travaux de cette thèse concernent l'amélioration du processus de prise de décision en unité de soins intensifs néonatals et des résultats néonatals à l'aide de méthodes non invasives et continues, basées sur le traitement des signaux physiologiques et des algorithmes d'apprentissage automatique, combinant des approches basées sur des modèles et des connaissances. Les principales contributions comprennent (Figure 1) :

1. Sur la base des travaux de notre équipe (*SEPIA* in *LTSI - INSERM U1099*), cette thèse a intégré et amélioré **une chaîne complète de traitement automatique des signaux**, qui sert d'outil fondamental tout au long de la thèse. Elle comprend plusieurs étapes clés : l'évaluation de la qualité et le traitement des signaux ECG, la détection du complexe QRS, l'extraction et la correction des intervalles RR, le débruitage des séries temporelles, l'analyse de la stationnarité et l'analyse de la variabilité de la fréquence cardiaque. Cette chaîne de traitement utilise comme entrée les signaux de surveillance cardiaque provenant de la pratique clinique (données de vie réelle), et produit des segments stationnaires optimaux et des paramètres utiles pour l'analyse ultérieure et le développement de modèles d'apprentissage automatique.
2. Concernant **la gestion de l'hyperbilirubinémie néonatale**, nous avons étudié deux aspects importants : (a) la caractérisation de la dynamique de la bilirubine sérique totale (BST) basée sur un modèle mathématique et (b) l'estimation non invasive de la BST basée sur les connaissances à l'aide de modèles d'apprentissage automatique à effets mixtes. Une contribution majeure dans cette application a été la proposition de modèles hybrides intégrant des modèles à effets mixtes, des modèles d'apprentissage automatique et des modèles physiologiques. Ensemble, ces études visent offrir une approche globale pour relever les défis cliniques de l'hyperbilirubinémie chez les prématurés et à ouvrir la voie à des interventions cliniques plus efficaces et spécifiques au patient.
   (a) **Caractérisation des dynamiques de la bilirubine sérique totale basée sur un modèle :**
   Nous avons proposé et validé un modèle de décroissance exponentielle spécifique à chaque patient pour caractériser la dynamique naturelle et à long terme des concentrations de bilirubine sérique totale chez les nourrissons prématurés nés entre 24 et 32 semaines de gestation. Grâce à un ajustement personnalisé, nous avons obtenu 72 modèles avec des paramètres spécifiques à chaque patient, optimisés en minimisant l'erreur entre les niveaux mesurés de BST et les résultats du modèle, à l'aide d'une méthode adaptative robuste aux moindres carrés. Le modèle proposé a démontré son efficacité et sa capacité à suivre de près les niveaux observés de BST pendant des périodes néonatales prolongées, avec une racine de l'erreur quadratique moyenne allant de 1,20 à 40,25 µmol/L, avec une médiane [écart interquartile] de 8,74 [4,89 ; 14,25] µmol/L. De plus, lorsque l'évolution de la bilirubine d'un patient diverge du modèle attendu de décroissance, comme indiqué par une augmentation de la racine de l'erreur quadratique moyenne, cela peut suggérer la survenue d'événements cliniques à haut risque tels qu'une entérocolite nécrosante et des niveaux élevés de protéine C réactive.

Cette association indique que les capacités du modèle dépassent la simple analyse descriptive et peuvent servir de nouveaux biomarqueurs potentiels pour la détection précoce de comorbidités pertinentes.

(b) **Estimation non invasive de la bilirubine totale basée sur des connaissances à l'aide de modèles d'apprentissage automatique à effets mixtes :**
Nous avons exploré des approches non invasives pour estimer les niveaux de TSB chez les nourrissons prématurés, avec ou sans hyperbilirubinémie, nés entre $24^{2/7}$ et $31^{6/7}$ semaines de gestation. Nous avons comparé différents estimateurs d'apprentissage automatique en intégrant des effets mixtes et des représentations de connaissances physiologiques. Par rapport à une forêt aléatoire standard, les modèles proposés de forêt aléatoire à effets mixtes modifiée (MERF) ont considérablement amélioré les accords d'estimation et réduit les biais proportionnels grâce à l'intégration explicite de connaissances physiologiques pertinentes. Bien que ces modèles nécessitent des données historiques spécifiques à chaque patient pour leur initialisation, ils montrent un potentiel clinique dans les USIN, où les données cliniques longitudinales sont fréquemment disponibles.

3. Concernant **la détection précoce de la septicémie néonatale tardive**, une première contribution a été **la formalisation de la chronologie clinique de la surveillance, la détection et la confirmation de la septicémie**, avec des constantes de temps estimées à partir de la littérature ou directement à partir de nos données CARESS-Premi. Cette formalisation facilite la représentation des effets causaux qui interviennent lors de la prise de décision en matière de détection précoce du sepsis et nous a permis de proposer une approche originale dans ce domaine. En effet, dans ce travail, nous avons évalué l'efficacité de la détection du risque de sepsis sans tenir compte de la suspicion clinique (c'est-à-dire avant le début du traitement et de l'intervention) lors de l'entraînement des modèles, ce qui n'est couramment pas le cas dans la littérature.

   Une deuxième contribution dans cet axe a été **la constitution d'une base de données longitudinale formellement annotée de la surveillance multiparamétrique des signaux en USIN pour la détection précoce du sepsis néonatal**. Nous avons acquis les signaux de surveillance continue et longitudinale ECG d'environ 450 prématurés et les avons segmentés en blocs de 6 heures. La chaîne de traitement des signaux proposée a été appliquée sur tous les segments de données, afin de dériver les caractéristiques HRV. Une stratégie formalisée d'étiquetage a été conçue pour générer des pseudo-étiquettes pour chaque segment, qui indiquent l'état du patient (septicémie, non-septicémie ou incertain). Cette base de données pourrait être utilisée dans d'autres recherches de notre équipe, en particulier pour l'exposition conjointe des signaux ECG et respiratoires.

   Une troisième contribution dans cet axe a été **la proposition et l'évaluation, sur la base de données susmentionnée, d'une série de modèles d'apprentissage automatique**, y com-

pris la régression logistique (LR), la forêt aléatoire (RF), l'eXtreme Gradient Boosting (XGBoost), le perceptron multicouche (MLP), ainsi que des réseaux de neurones convolutifs peu profonds (shallow CNN). Selon leur degré d'indépendance temporelle, les détecteurs développés ont été catégorisés en détecteurs instantanés (indépendants du temps) et détecteurs dépendants du temps. La meilleure performance mesurée parmi les détecteurs instantanés, basée sur l'analyse des courbes ROC et PRC, a été obtenue par un classifieur RF, avec des valeurs respectives de 0,727$\pm$0,060 et 0,352$\pm$0,099. Les shallow CNN ont fourni les meilleures performances en tant que détecteurs à court terme dépendant du temps, avec des courbes ROC de 0,737$\pm$0,027 (en utilisant trois segments successifs) et de 0,749$\pm$0,045 (en utilisant six segments successifs). Une analyse de sensibilité a été réalisée pour améliorer l'interprétabilité des modèles. Il convient de noter que nos résultats ne sont pas comparables aux performances élevées rapportées dans certains articles de la littérature (AUC > 0,80) et, plus important encore, nous affirmons qu'il est impossible de faire des comparaisons directes entre les études étant donné l'énorme hétérogénéité et la grande variabilité de nombreux aspects tels que les définitions de la septicémie néonatale, les stratégies d'annotation, etc. Nous proposons une critique sur la manière dont la plupart de ces travaux de la littérature ont été construits et évalués, et nous proposons une discussion sur les défis liés à une évaluation correcte des performances dans ce contexte. Deux limitations particulières sont mises en évidence : *i)* l'absence d'une définition précise et normalisée de l'instant de début d'un épisode de sepsis et *ii)* le potentiel très élevé de sous-estimation des faux positifs dans ces travaux. Nous considérons que l'entrainement de modèles ML supervisés dans la plupart de ces publications souffre de sources significatives de biais, ce qui peut expliquer leur manque d'applicabilité clinique et de généralisation. Un effort significatif doit être fait dans ce domaine afin d'envisager des applications cliniques utiles à l'avenir. Cette discussion, ainsi que les efforts réalisés dans ce travail pour l'annotation et le traitement formels et temporels de la base de données CARESS-Premi sont pour nous une contribution supplémentaire de ce doctorat.

4. Enfin, la dernière contribution de cette thèse concerne **le développement et le déploiement d'un prototype avancé de CDSS**, dédié à la détection précoce, le diagnostic et l'intervention dans un contexte clinique. Nous avons conçu, mis en oeuvre, déployé et évalué techniquement un CDSS qui intègre des chaînes de traitement du signal en quasi-temps réel et des modèles d'intelligence artificielle en inférence, dans l'USIN du CHU de Rennes. Ce système comprend la transmission de données, la pseudonymisation, la fusion de données, le traitement des signaux et l'application des modèles en inférence. Au cours des six premiers mois de déploiement depuis janvier 2023, le service a reçu en continu les données de monitoring de 138 nouveau-nés, traitant les données en direct et générant des estimations du niveau de bilirubine à une résolution temporelle de 15 minutes. Malgré certaines limitations du modèle d'inférence actuel pour l'estimation de la bilirubine, qui constituait le premier cas d'usage intégré au système *on-the-edge*, une évaluation quantitative des performances

techniques du système en termes de stabilité et de consommation des ressources a été réalisée, et une configuration du système robuste et satisfaisante a été obtenue. En outre, un rapport de cas clinique spécifique, basé sur les résultats intermédiaires dérivés du système en quasi-temps réel, apporte des indications prometteuses quant à l'utilité et à la faisabilité du système. À notre connaissance, il s'agit de la première proposition d'un CDSS complet, multi-sources, quasi-temps-réel et *on-the-edge*, déployé en vie-réelle pour la détection précoce d'événements à haut risque spécifiques à chaque patient. Cette démonstration de faisabilité constitue un premier pas solide vers la mise en place d'un système de détection précoce des événements à haut risque, en exploitant les propriétés dynamiques des données longitudinales multivariées et multi-sources sur la plateforme proposée.

En conclusion, en s'appuyant sur le traitement des signaux physiologiques et les techniques d'apprentissage automatique, cette thèse de doctorat a proposé différentes approches basées sur des connaissances et des modèles pour améliorer le diagnostic et la gestion des conditions néonatales critiques telles que l'hyperbilirubinémie et le sepsis tardif. De plus, un CDSS *on-the-edge*, en tant que preuve de concept, a été déployé et évalué de manière préliminaire dans un environnement réel d'USIN, démontrant l'efficacité de la pipeline et des modèles proposés en temps réel. Dans l'ensemble, les explorations de cette thèse montrent un potentiel prometteur pour l'optimisation de la surveillance en USIN et l'amélioration de la détection précoce des événements à haut risque. Ces avancées pourraient grandement renforcer la prise en charge et les résultats des nouveau-nés prématurés.

BIBLIOGRAPHY

[1] Definitions, WHO Recommended, "Terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths," Acta Obstet Gynecol Scand, vol. 56, no. 3, pp. 247–53, 1977.

[2] World Health Organization, Born too soon : decade of action on preterm birth. World Health Organization, 2023.

[3] H. Cinelli, N. Lelong, and C. Le Ray, "Enquête nationale périnatale. rapport 2021," 2022.

[4] J. Perin, A. Mulick, D. Yeung, F. Villavicencio, G. Lopez, K. L. Strong, D. Prieto-Merino, S. Cousens, R. E. Black, and L. Liu, "Global, regional, and national causes of under-5 mortality in 2000–19 : an updated systematic analysis with implications for the sustainable development goals," The Lancet Child & Adolescent Health, vol. 6, no. 2, pp. 106–115, 2022.

[5] R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, "Epidemiology and causes of preterm birth," The lancet, vol. 371, no. 9606, pp. 75–84, 2008.

[6] A. J. Masino, M. C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C. P. Bonafide, F. Balamuth, M. Schmatz, and R. W. Grundmeier, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," PloS one, vol. 14, no. 2, p. e0212665, 2019.

[7] E. Persad, K. Jost, A. Honoré, D. Forsberg, K. Coste, H. Olsson, S. Rautiainen, and E. Herlenius, "Neonatal sepsis prediction through clinical decision support algorithms : a systematic review," Acta Paediatrica, vol. 110, no. 12, pp. 3201–3226, 2021.

[8] J. S. Gilchrist, "Clinical decision support system using real-time data analysis for a neonatal intensive care unit," Ph.D. dissertation, Carleton University, 2012.

[9] A. Rao and J. Palma, "Clinical decision support in the neonatal ICU," in Seminars in Fetal and Neonatal Medicine, vol. 27, no. 5. Elsevier, 2022, p. 101332.

[10] M. Altuve, "Détection multivariée des épisodes d'apnée-bradycardie chez le prématuré par modèles semi-markovien cachés," Thèse de Doctorat, Université de Rennes 1, Jan. 2011.

[11] Y. Wang, "Heart rate variability and respiration signals as late onset sepsis diagnostic tools in neonatal intensive care units," Thèse de Doctorat, Université de Rennes 1, Dec. 2013.

[12] M. Calvo Gonźalez, "Analysis of the cardiovascular response to autonomic nervous system modulation in Brugada syndrome patients," Thèse de Doctorat, Université de Rennes 1, Nov. 2017.

[13] P. T. T. Nguyen, "Aide au diagnostic précoce par analyse de la variabilité cardiaque et respiratoire chez le nouveau-né," Thèse de Doctorat, Université de Rennes 1, Dec. 2019.

[14] S. Cabon, "Monitoring of premature newborns by video and audio analyses," Ph.D. dissertation, Université de Rennes, Jul. 2019.

[15] G. Guerrero, "Analyse à base de modèles des interactions cardiorespiratoires chez l'adulte et chez le nouveau-né," Thèse de Doctorat, Université de Rennes 1, Sep. 2020.

[16] C. S. Leon Borrego, "Apprentissage automatique pour la prédiction de l'infection et de la maturation chez le grand prématuré en associant les variabilités cardiaques et respiratoires," Thèse de Doctorat, Université de Rennes 1, Jul. 2021.

[17] B. Met-Montot, "Detection and characterization of vocalizations in preterm newborns," Thèse de Doctorat, Université de Rennes, Jun. 2022.

[18] O. Duport, "Analyse à base de modèles des interactions cardio-respiratoires chez le nouveau-né," Thèse de Doctorat, Université de Rennes, Jun. 2023.

[19] E. H. Shortliffe and M. J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," Jama, vol. 320, no. 21, pp. 2199–2200, 2018.

[20] S. Ramgopal, L. N. Sanchez-Pinto, C. M. Horvat, M. S. Carroll, Y. Luo, and T. A. Florin, "Artificial intelligence-based clinical decision support in pediatrics," Pediatric research, vol. 93, no. 2, pp. 334–341, 2023.

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

[22] A. R. Kemper, T. B. Newman, J. L. Slaughter, M. J. Maisels, J. F. Watchko, S. M. Downs, R. W. Grout, D. G. Bundy, A. R. Stark, D. L. Bogen et al., "Clinical practice guideline revision : management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation," Pediatrics, vol. 150, no. 3, 2022.

[23] B. J. Stoll, N. Hansen, A. A. Fanaroff, L. L. Wright, W. A. Carlo, R. A. Ehrenkranz, J. A. Lemons, E. F. Donovan, A. R. Stark, J. E. Tyson et al., "Late-onset sepsis in very low birth weight neonates : the experience of the NICHD Neonatal Research Network," Pediatrics, vol. 110, no. 2, pp. 285–291, 2002.

[24] S. A. Coggins and K. Glaser, "Updates in late-onset sepsis : risk assessment, therapy, and outcomes," Neoreviews, vol. 23, no. 11, pp. 738–755, 2022.

[25] R. Hayes, J. Hartnett, G. Semova, C. Murray, K. Murphy, L. Carroll, H. Plapp, L. Hession, J. OToole, D. McCollum et al., "Neonatal sepsis definitions from randomised clinical trials," Pediatric research, vol. 93, no. 5, pp. 1141–1148, 2023.

[26] A. L. Shane, P. J. Sánchez, and B. J. Stoll, "Neonatal sepsis," The lancet, vol. 390, no. 10104, pp. 1770–1780, 2017.

[27] Z. Uhrikova, M. Zibolen, K. Javorka, L. Chladekova, and M. Javorka, "Hyperbilirubinemia and phototherapy in newborns : Effects on cardiac autonomic control," Early Human Development, vol. 91, no. 6, pp. 351–356, 2015.

[28] M.-L. Specq, M. Bourgoin-Heck, N. Samson, F. Corbin, C. Gestreau, M. Richer, H. Kadhim, and J.-P. Praud, "Moderate hyperbilirubinemia alters neonatal cardiorespiratory control and induces inflammation in the nucleus tractus solitarius," Frontiers in physiology, vol. 7, p. 437, 2016.

[29] R. Özdemir, Ö. Olukman, C. Karadeniz, K. Çelik, N. Katipoğlu, M. Muhtar Yılmazer, Ş. Çalkavur, T. Meşe, and S. Arslanoğlu, "Effect of unconjugated hyperbilirubinemia on neonatal autonomic functions : Evaluation by heart rate variability," The Journal of Maternal-Fetal & Neonatal Medicine : The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians, vol. 31, no. 20, pp. 2763–2769, 2018.

[30] S. Al-Omar, V. Le Rolle, N. Samson, M.-L. Specq, M. Bourgoin-Heck, N. Costet, G. Carrault, and J.-P. Praud, "Influence of moderate hyperbilirubinemia on cardiorespiratory control in preterm lambs," Frontiers in physiology, vol. 10, p. 468, 2019.

[31] J. R. Moorman, D. E. Lake, and M. P. Griffin, "Heart rate characteristics monitoring for neonatal sepsis," IEEE Transactions on Biomedical Engineering, vol. 53, no. 1, pp. 126–132, 2005.

[32] A. Beuchée, G. Carrault, J. Y. Bansard, E. Boutaric, P. Bétrémieux, and P. Pladys, "Uncorrelated randomness of the heart rate is associated with sepsis in sick premature infants," Neonatology, vol. 96, no. 2, pp. 109–114, 2009.

[33] K. D. Fairchild and T. M. O'Shea, "Heart rate characteristics : physiomarkers for detection of late-onset neonatal sepsis," Clinics in perinatology, vol. 37, no. 3, pp. 581–598, 2010.

[34] L. Rio, A.-S. Ramelet, P. Ballabeni, C. Stadelmann, S. Asner, and E. Giannoni, "Monitoring of heart rate characteristics to detect neonatal sepsis," Pediatric Research, vol. 92, no. 4, pp. 1070–1074, 2022.

[35] Z. Peng, G. Varisco, R.-H. Liang, D. Kommers, W. Cottaar, P. Andriessen, C. van Pul, and X. Long, "DeepLOS : Deep learning for late-onset sepsis prediction in preterm infants using heart rate variability," Smart Health, vol. 26, p. 100335, 2022.

Table of Contents

Abbreviations and Acronyms

CHU Rennes *University Hospital Center of Rennes*. 51, 120, 227, 229, 230, 237, 240, 242, 243, 250

LTSI - INSERM U1099 *Laboratoire Traitement du Signal et de l'Image (Image and Signal Processing Laboratory)*. 52, 53, 228, 237, 242, 247

SEPIA *Stimulation thÉrapeutique et monitoring Personnalisés pour l'Insuffisance cardiaque et les Apnées-bradycardies (LTSI team)*. 52, 53, 147, 247

AAP American Academy of Pediatrics. 8, 11, 14, 15, 18, 30, 114, 115

ABE Acute bilirubin encephalopathy. 13, 15

AI Artificial Intelligence. 3, 27–33, 85, 86, 97, 98, 213, 215, 216, 227, 244

ANN Artificial neural networks. 31, 90

ANS Autonomic nervous system. 70, 74, 75, 144, 147, 172

ASCENT Anonymised System for Clinical experimENTation. 53

AUPRC Area under the precision-recall curve. 96, 193, 194, 196, 214, 249

AUROC Area under the receiver operating characteristic curve. 30, 31, 96, 193, 194, 196, 213–215, 249

BAcc Balanced accuracy. 95, 193, 194, 196

BIND Bilirubin-induced neurologic dysfunction. 13, 114

BPD Bronchopulmonary dysplasia. 9, 11

BW Birth weight. 20, 120, 136

CA Corrected age. 8

CART Classification and negression tree. 87, 88, 91

CBE Chronic bilirubin encephalopathy. 13

CDSS Clinical decision support systems. 1–3, 26–33, 35, 51, 52, 174, 227–230, 241–244, 250, 251

CNN Convolutional neural networks. 31, 90, 91, 188, 191, 192, 196, 210, 214, 218, 249

CNS Central nervous system. 13

CRP C-reactive protein. 24, 55, 120, 123, 129–131, 137, 138, 172, 174, 175, 180, 181, 184, 211, 212, 242, 265–272

DFA Detrended fluctuation analysis. 74

DT Decision Tree. 31, 87–89

DWC Data Warehouse Connect. 228–232, 236, 238, 240

ECG Electrocardiogram. 30, 52, 53, 56, 59–64, 66, 68, 144–146, 175, 178–180, 186, 192, 210, 212–214, 216, 217, 236, 243, 247, 249

EE Elementary Effects. 102, 103, 255

EEG Electroencephalogram. 29, 31

EHR Electronic health records. 26–31, 34, 180, 212, 213, 215, 230, 236, 243

ELBW Extremely low birth weight. 23

EM Expectation-Maximization. 92, 149, 151, 164

EOS Early-onset sepsis. 23, 24, 30

ET Exchange transfusion. 15, 18, 120

FNR False negative rate. 194, 214

FPR False positive rate. 95, 96, 194, 214

G6PD Glucose-6-phosphate dehydrogenase. 13, 16, 17

GA Gestational age. 8, 14, 18, 20, 21, 56, 114, 120, 121, 130, 132, 136, 152, 180, 189, 192, 201, 205, 210, 230, 236, 237, 265–272

GBDT Gradient Boosting Decision Tree. 89

GLL Generalized log-likelihood. 92, 150, 154, 156

HDF5 Hierarchical data format. 56, 145, 178

HIE Hypoxic-ischaemic encephalopathy injury. 9, 31

HIS Hospital Information System. 26–28, 31, 227, 230, 240, 243, 251

HR Heart rate. 63, 71–73, 170, 171

HRC Heart rate characteristics. 170, 171

HRV Heart rate variability. 2, 3, 30, 59, 64, 70–72, 74, 144, 147–149, 151, 152, 162, 163, 169–172, 174–176, 178–180, 188–190, 192, 193, 196, 201, 204, 205, 210, 213, 214, 216–218, 234, 236, 240, 242, 249, 250

ICU Intensive care units. 228

IQR Interquartile range. 21, 23, 25, 82, 83, 125, 129, 138, 237, 248

IVH Intraventricular hemorrhage. 52, 55, 131

LMM Linear mixed-effects model. 91, 92

LoA Limits of agreement. 80–82, 151, 152, 156, 158

LOS Late-onset sepsis. 3, 23, 24, 30, 34, 169–172, 174, 175, 177, 186, 191–193, 201, 210–213, 217, 218, 240, 247, 249

LR Logistic Regression. 87, 188, 192, 196, 213, 249

MAE Mean absolute error. 97

MDI Mean decrease in impurity. 100

MERF Mixed-Effects Random Forest. 91–93, 143, 149–151, 159, 162–164, 248, 251

MERT Mixed-Effects Regression Tree. 91, 92

MFPD Multi-Feature Probabilistic Detector. 64

ML Machine learning. 2, 27, 29, 30, 33, 34, 85, 86, 88, 93, 97, 99, 101, 102, 159, 164, 175, 176, 187, 192–194, 210–215, 218, 230, 236, 242–244, 247–250

MLP Multilayer perceptron. 90, 188, 192, 194, 196, 249

mMERF Modified Mixed-Effects Random Forest. 150, 162

MSE Mean squared error. 97

NaN Not-a-Number. 62, 63, 67, 68

NB Naïve Bayes. 86, 188, 192, 194, 213

NEC Necrotizing enterocolitis. 9, 11, 114, 120, 123, 131, 137

NHB Neonatal hyperbilirubinemia. 3, 11, 14, 15, 113–115, 143, 247, 248

NICE National Institute for Health and Care Excellence. 11, 14, 15, 18–21, 25, 120, 124, 130–132, 136, 265–272

NICU Neonatal intensive care units. 1–3, 7, 10, 11, 22, 25, 26, 28–35, 51, 53, 57, 60, 104, 113, 137, 138, 143–145, 165, 169–172, 174, 175, 210, 213, 217, 227, 229, 237, 240, 242–244, 247–252

OLS Ordinary least squares. 77, 121

OOB Out-of-bag. 98, 99

OTE On-the-edge. 164

PCA Principal component analysis. 83

PDA Patent ductus arteriosus. 9, 30, 131

PIC iX Patient Information Center iX. 228, 229

PMA Postmenstrual age. 8, 149–151, 156, 162, 230

PNA Postnatal age. 8, 56, 120–125, 127, 128, 130, 132, 133, 149–152, 156, 162, 180, 189, 192, 201, 210, 230, 236, 256

PPG Photoplethysmography. 52

PPV Positive predictive value. 95

PRC Precision-recall curve. 96, 194

PSD Power spectral density. 72, 73

PT Phototherapy. 15, 120, 131

R^2 R-squared. 97

RDS Respiratory distress syndrome. 9

RF Random Forest. 87–89, 91–93, 98–101, 143, 148–151, 159, 162–164, 188, 192, 194, 196, 201, 210, 214, 237, 248, 249

RMSE Root-mean-square error. 97, 123, 125, 129, 131, 132, 136, 138, 151, 156, 248, 258, 265–272

ROC Receiver operating characteristic curve. 96, 194, 214

ROR Rate of rise. 114, 116

RRI RR intervals. 63, 64, 68

RT Regression Tree. 91, 92

SD Standard deviation. 80, 81, 125

SEM Standard error of the mean. 206–209

SNR Signal-to-noise ratio. 60, 62, 64, 180, 237

SSR Sum of squared residuals. 77

SVM Support vector machine. 31, 90

TcB Transcutaneous bilirubin. 14, 114, 144, 163

TNR True negative rate. 95, 193, 196

TPR True positive rate. 95, 96, 193, 194, 196

TSB Total serum bilirubin. 13–15, 18, 20–22, 113–125, 127–133, 136–138, 143–145, 147–152, 156, 159, 162, 163, 165, 248, 251

UI User interface. 27, 32

VI Variable importance. 98, 99

VLBW Very low birth weight. 23, 115, 171

VM Virtual machine. 230, 234, 237, 240, 242

XGBoost eXtreme Gradient Boosting. 87, 89, 188, 192, 213, 249

Introduction

Preterm infants, defined as neonates born alive before 37 complete weeks of gestation [1], account for a substantial portion of the global neonatal population, with approximately 15 million preterm births occurring annually—equivalent to 1 in 10 newborns worldwide [2]. In France, 122 preterm births are recorded daily, amounting to 44,500 annually, representing 7.0% of all births [3]. Due to their early arrival, these newborns often face incomplete development of vital organ systems, rendering them particularly immature and susceptible to a range of complications, including cardio-respiratory, immunological, neurological, and digestive conditions. The **complications associated with preterm birth** remain the leading cause of under-5 child mortality [2, 4] and contribute to more than half of long-term morbidity cases [5].

Neonatal care, particularly in the **Neonatal intensive care units (NICU)**, is crucial for the survival and development of preterm and critically ill newborns. Given their vulnerability, NICU provide these fragile newborns with continuous, specialized medical care through life-support systems, advanced monitoring technologies and tailored clinical interventions. However, despite advances in neonatal medicine and care practices, managing high-risk conditions in NICU remains a complex challenge. Two of the most prevalent and difficult conditions encountered in these settings are **neonatal hyperbilirubinemia** (jaundice) [6, 7] and **neonatal sepsis** [8–11]. These complications are not only consistently associated with prolonged hospital stays but also significantly impact the overall outcomes of preterm infants.

Improving health outcomes for preterm infants relies on the **ability to diagnose and intervene in critical conditions as early as possible**. Early detection of adverse events, particularly through non-invasive methods, remains a challenge but has become a key trend in neonatal care. In recent years, **Clinical decision support systems (CDSS)**—leveraging **big data, evidence-based medicine, and advanced analytical techniques** such as artificial intelligence—have emerged as promising tools to enhance early detection and assist clinicians in making more informed, timely decisions [12–15]. However, integrating these systems into clinical workflows presents significant challenges, including complexities in data acquisition and processing, the need for model explainability, and the difficulties of real-world implementation [16, 17]. Addressing these challenges is crucial for the successful **development and deployment of CDSS in NICU**, where massive amounts of monitoring data can be transformed into actionable insights, ultimately improving both short- and long-term outcomes for preterm infants.

Particularly, this dissertation is interested in a set of **major methodological challenges** inherent to working with Machine learning (ML) models on NICU neonatal monitoring data, which are longitudinal, continuous, time-dependent, and often noisy. Key challenges often overseen in this context include:

- Temporal dependencies & time-varying covariates: Longitudinal data often exhibits temporal dependencies between observations, and covariates (predictor variables) can change over time, making it challenging for traditional feature engineering and machine learning algorithms, that assume independence.
- Missing values: Longitudinal data often contain missing values due to measurement errors or non-response rates, which can lead to biased estimates if not handled properly.
- Temporal granularity & non-stationarity: Data may be collected at varying time scales and exhibit time-series properties like seasonality or trends, while patterns and distributions can shift over time, requiring adaptive modeling.
- Model complexity & scalability: The complexity of longitudinal data may increase model complexity, raising the risk of overfitting and making it computationally expensive, thus demanding efficient regularization techniques and scalable models.
- Model interpretability: Understanding how models use temporal information and how this affects predictions is critical to ensuring they are meaningful and actionable.

In this context, the overall objective of this Ph.D. dissertation is **to develop advanced data processing techniques and explainable machine learning models to improve the diagnosis and management of two critical clinical challenges in the NICU**: neonatal hyperbilirubinemia and late-onset sepsis. This research aims **to facilitate decision-making in NICU by leveraging non-invasive, continuous, and near real-time monitoring systems, providing clinicians with actionable insights to improve neonatal outcomes**. Specifically, the objectives are as follows:

- **To improve diagnosis and therapy optimization** for preterm infants by analyzing dynamic features such as Heart rate variability (HRV) extracted from physiological signals (e.g., cardio-respiratory recordings), using advanced and specific **data processing** methods and **machine learning** algorithms.
- **To deepen the understanding of physiological patterns in preterm infants** by analyzing longitudinal data and correlating it with clinical outcomes using both model-based and knowledge-based approaches.
- **To integrate the proposed models and algorithms into continuous monitoring systems** and Clinical decision support systems, enabling near real-time inference to assist clinicians by providing timely, data-driven insights.

This dissertation is organized as follows (Figure 2):

Chapter 1 provides the clinical context of preterm birth, focusing on the challenges in managing Neonatal hyperbilirubinemia (NHB) and Late-onset sepsis (LOS) in NICU. It also introduces the role of Clinical decision support systems (CDSS) in enhancing neonatal care through brief reviews of existing CDSS technologies, standards for AI-based deployment, and the challenges involved in integrating these systems into NICU settings.

Chapter 2 outlines the key methodologies and tools employed throughout the dissertation, detailing the CARESS-Premi project and its data acquisition process, the proposed data processing pipeline, statistical techniques, machine learning algorithms, and explainability analyses. The methodological framework serves as the foundation for the subsequent studies.

Chapter 3 is dedicated to the first challenge of managing neonatal hyperbilirubinemia. It focuses on the **model-based characterization of natural bilirubin dynamics in preterm infants** during the postnatal period by developing and validating a patient-specific exponential decay model. It also explores the potential of the model parameters as biomarkers for detecting associated morbidities.

Chapter 4 extends the discussion on NHB management by proposing **a knowledge-based, non-invasive approach to estimate bilirubin levels** using modified mixed-effects random forests, building on the physiological insights from the exponential decay model proposed in Chapter 3 to improve accuracy.

Chapter 5 shifts focus to the challenge of **early detection of neonatal Late-onset sepsis (LOS) in preterm infants using HRV data derived from real-life monitoring data**. A set of machine learning algorithms incorporating different HRV variants are proposed and compared to detect septic events before clinical suspicion emerges.

Chapter 6 presents the design and implementation of a quasi-real-time CDSS, integrating signal processing and machine learning models on the edge in the scope of NICU. It shows the system's technical performance and clinical feasibility, with preliminary results from first use cases.

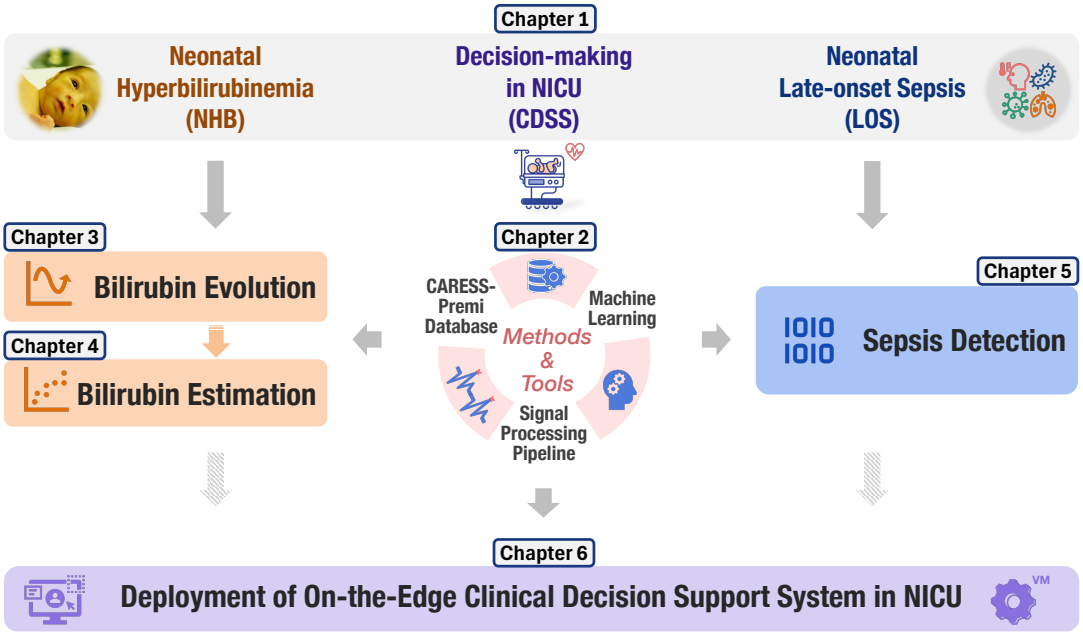

Figure 2: Scope and framework of the dissertation.

BIBLIOGRAPHY

[1] Definitions, WHO Recommended, "Terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths," Acta Obstet Gynecol Scand, vol. 56, no. 3, pp. 247–53, 1977.

[2] World Health Organization, Born too soon: decade of action on preterm birth. World Health Organization, 2023.

[3] H. Cinelli, N. Lelong, and C. Le Ray, "Enquête nationale périnatale. rapport 2021," 2022.

[4] J. Perin, A. Mulick, D. Yeung, F. Villavicencio, G. Lopez, K. L. Strong, D. Prieto-Merino, S. Cousens, R. E. Black, and L. Liu, "Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the sustainable development goals," The Lancet Child & Adolescent Health, vol. 6, no. 2, pp. 106–115, 2022.

[5] R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, "Epidemiology and causes of preterm birth," The lancet, vol. 371, no. 9606, pp. 75–84, 2008.

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

[7] A. R. Kemper, T. B. Newman, J. L. Slaughter, M. J. Maisels, J. F. Watchko, S. M. Downs, R. W. Grout, D. G. Bundy, A. R. Stark, D. L. Bogen et al., "Clinical practice guideline revision: management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation," Pediatrics, vol. 150, no. 3, 2022.

[8] B. J. Stoll, N. Hansen, A. A. Fanaroff, L. L. Wright, W. A. Carlo, R. A. Ehrenkranz, J. A. Lemons, E. F. Donovan, A. R. Stark, J. E. Tyson et al., "Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network," Pediatrics, vol. 110, no. 2, pp. 285–291, 2002.

[9] S. A. Coggins and K. Glaser, "Updates in late-onset sepsis: risk assessment, therapy, and outcomes," Neoreviews, vol. 23, no. 11, pp. 738–755, 2022.

[10] R. Hayes, J. Hartnett, G. Semova, C. Murray, K. Murphy, L. Carroll, H. Plapp, L. Hession, J. O' Toole, D. McCollum et al., "Neonatal sepsis definitions from randomised clinical trials," Pediatric research, vol. 93, no. 5, pp. 1141–1148, 2023.

[11] A. L. Shane, P. J. Sánchez, and B. J. Stoll, "Neonatal sepsis," The lancet, vol. 390, no. 10104, pp. 1770–1780, 2017.

[12] A. J. Masino, M. C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C. P. Bonafide, F. Balamuth, M. Schmatz, and R. W. Grundmeier, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," PloS one, vol. 14, no. 2, p. e0212665, 2019.

[13] E. Persad, K. Jost, A. Honoré, D. Forsberg, K. Coste, H. Olsson, S. Rautiainen, and E. Herlenius, "Neonatal sepsis prediction through clinical decision support algorithms: a systematic review," Acta Paediatrica, vol. 110, no. 12, pp. 3201–3226, 2021.

[14] J. S. Gilchrist, "Clinical decision support system using real-time data analysis for a neonatal intensive care unit," Ph.D. dissertation, Carleton University, 2012.

[15] A. Rao and J. Palma, "Clinical decision support in the neonatal ICU," Seminars in Fetal and Neonatal Medicine, vol. 27, no. 5, p. 101332, 2022.

[16] E. H. Shortliffe and M. J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," Jama, vol. 320, no. 21, pp. 2199–2200, 2018.

[17] S. Ramgopal, L. N. Sanchez-Pinto, C. M. Horvat, M. S. Carroll, Y. Luo, and T. A. Florin, "Artificial intelligence-based clinical decision support in pediatrics," Pediatric research, vol. 93, no. 2, pp. 334–341, 2023.

CHAPTER

1 Context

The opening chapter of this dissertation provides a comprehensive context and interest in the management of preterm infants in Neonatal intensive care units (NICU) settings, with a focus on the high-risk events they commonly encounter. An overview of preterm birth in terms of definitions and associated complications will be first presented, followed by a brief description of the NICU, where preterm and critically ill newborns receive specialized medical care. Next, two primary challenges in NICU as the main interests of this dissertation, neonatal hyperbilirubinemia and neonatal sepsis, are introduced. Finally, the chapter explores the integration of clinical decision support systems in NICU, discussing both the current challenges and the promising opportunities in its implementation for improving neonatal care.

1.1 Preterm Birth

1.1.1 Definition

Pregnancy refers to the period during which a fetus develops within a woman's uterus, measured by the time elapsing from the first day of the last menstrual period to the day of delivery, known as gestation. A typical human pregnancy lasts approximately 40 weeks (9.2 months). According to the latest guidelines, health care providers define pregnancy as full-term when it lasts between 39 weeks and 40 weeks and 6 days [1]. Infants born during this duration are considered full-term infants.

Preterm birth, also known as premature birth, occurs before 37 completed weeks (259 days) of gestation, and accordingly, all babies born alive in this case are considered preterm neonates [2]. According to a 2023 report released by the United Nations agencies and partners, a cumulative total of 152 million vulnerable babies were born prematurely globally in the past decade from 2010 to 2020. In 2020 alone, an estimated 13.4 million live births were preterm worldwide, equivalent to around 1 in 10 babies being born prematurely [3].

In China, over 7.5 million preterm births were recorded in 2020, making it the fourth-highest country in terms of preterm births, following India, Pakistan, and Nigeria, with a preterm rate of 6.1% [4].

In France, according to the latest statistics from DREES (Direction de la Recherche, des Études, de l'Evaluation et des Statistiques du Ministère de la Santé) [5], 122 preterm births occurred daily

in 2021, amounting to 44,500 preterm births for the year. The premature birth rate stood at 7.0%, with 11.0% of these classified as "small-for-gestational-age (less than the 10th percentile)" [1] [2].

Most preterm births are caused by spontaneous preterm labor, while some are medically indicated to be delivered by induced labor or cesarean section. However, the exact cause of preterm birth is difficult to identify and it may result from multi-fold factors at the same time as labor is a complex process [7, 8].

Standard age terminologies to describe the length of gestation and age in neonates during the perinatal period are defined [9] (Figure 1.1):

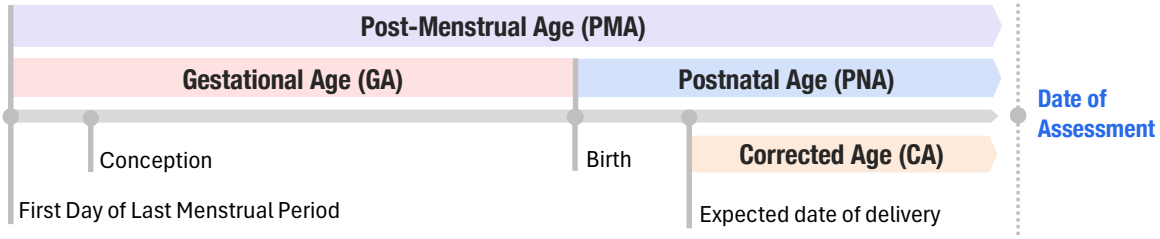

Figure 1.1: Age terminology during the perinatal period according to the American Academy of Pediatrics definitions.
(*Adapted from [9].*)

- **Gestational age (GA)**: the time elapsed between the first day of the last menstrual period and the day of delivery. Gestational age is usually expressed in completed weeks.
- **Postnatal age (PNA)**: or "chronological age", refers to the time elapsed since birth. It is usually written in days, weeks, months or years.
- **Postmenstrual age (PMA)**: the time elapsed between the first day of the last menstrual period and birth (gestational age) plus the time after birth (post-natal age). It is usually described in the number of weeks.
- **Corrected age (CA)**: or "adjusted age", the time elapsed between the expected date of birth and the date of assessment. It is usually described in weeks and months and is mainly used for premature infants under 3 years old.

1.1.2 Preterm birth associated mortality and morbidity

Preterm birth complications remain the leading driver of under-5 child mortality [3, 10] and account for more than half the long-term morbidity [11]. In 2021, an estimated 2.3 million neonatal deaths occurred worldwide (18 deaths per 1,000 live births), with approximately 1 million (0.9 million) newborns dying due to direct complications from preterm birth, and millions more survivors with disabilities that follow them and their families throughout their lives [3, 12].

1. Denominator: number of live births of singletons and twins
2. EPOPé curve [6], adjusted for gestational age and sex

Newborns born preterm face substantially higher risks of adverse outcomes compared to those born at term. The risks of mortality and morbidity increase with the degree of prematurity [4]. Preterm births are typically categorized into three subgroups based on accurate gestational age (Figure 1.2). In general, the more immature the infant, the greater the need for life support. However, it is essential to emphasize that all newborns, regardless of gestational age, should receive at least essential newborn care (level-I care) [3, 13]. More information on the levels of neonatal care is in [14].

- **Moderate to late preterm** refers to infants born between 32 and less than 37 weeks of gestation, accounting for approximately 85% of all preterm births. These infants are more fragile than full-term babies, in addition to level-I care, they may also require additional support or special newborn care (level-II care).
- **Very preterm** refers to infants born between 28 and less than 32 weeks of gestation and represents 10.5% of preterm births. These infants must benefit from additional support, typically provided through level-II care.
- **Extremely preterm** refers to infants born before 28 weeks of gestation. Though they represent only 4.5% of the prematurity, these infants are highly vulnerable and should be immediately transferred to intensive newborn care (level-III care) after birth.

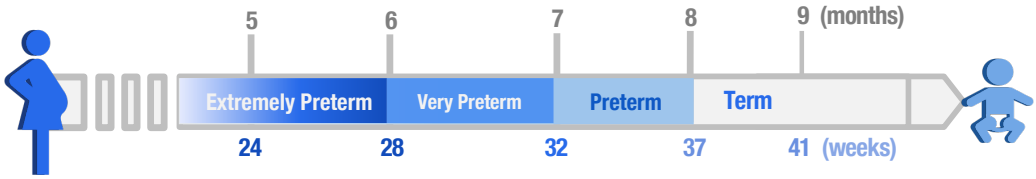

Figure 1.2: Sub-categories of preterm birth based on gestational age.

As preterm birth interrupts the newborn's development in utero, a range of complications arises from immature organ systems that are not yet ready to support life in the extra-uterine environment. This reflects the fragility and immaturity of critical systems, including the brain and brainstem, lungs, immune system, kidneys, skin, eyes and gastrointestinal system [15]. Consequently, preterm infants are exposed to severe conditions such as feeding difficulties, neonatal apnea, Respiratory distress syndrome (RDS), Bronchopulmonary dysplasia (BPD), Patent ductus arteriosus (PDA), Necrotizing enterocolitis (NEC), Hypoxic-ischaemic encephalopathy injury (HIE), hyperbilirubinemia (jaundice), kernicterus, and neonatal sepsis. Furthermore, this population is also at elevated risk of diverse long-term developmental challenges, including cerebral palsy, intellectual disabilities, epilepsy, visual or hearing impairments, and disorders in terms of psychological development, behavior and emotion.

1.2 Neonatal Intensive Care Units (NICU)

France was a forerunner in assisting premature infants in history, partly driven by its concerns over a falling birth rate [16]. By the 1970s, neonatal intensive care units flourished and became an established part of hospitals in developed countries, which helped to drastically decrease neonatal mortality and notably improve the quality of perinatal and neonatal care. Nowadays, as a necessity of level-III care, NICU are critical environments equipped with advanced technology, instruments and trained healthcare professionals. The NICU are designed to provide not only optimum around-the-clock care but also specialized life support and monitoring systems for the most vulnerable patients: preterm and critically ill infants.

The length of hospitalization of infants in the NICU depends on individual variability of health conditions and physiological competencies. As suggested in [17], thorough preparation and discharge planning is the process of working with a family to help them successfully transition from the NICU to home. Among these, babies must achieve some physiological milestones to demonstrate functional maturation before being discharged home, including [18, 19]:

- Antibiotics: to be off antibiotics and free from signs of infection;
- Feeding: achieve safe oral feeding to keep blood sugar level normal and support weight gain;
- Thermoregulation: be able to regulate their body temperature independently within a normal range;
- Events: to be free from apnea and/or bradycardic events during their "car seat test" to demonstrate cardio-respiratory stability;
- Respiratory control: be able to breathe independently and sufficiently and maintain respiratory stability.

Various medical supports are provided in the NICU in order to assist the tiniest newborns in attaining physiological maturity and meeting these milestones.

Incubators, where infants are placed, serve as the main elements in the NICU. These small and self-contained beds are enclosed by transparent, hard plastic shields, providing a specialized and controlled environment to maintain the infants at optimal temperature and humidity levels. Besides, incubators could serve a dual role in minimizing the risk of infections (cleaning with disinfectant wipes [20]) and reducing exposure to excessive noise (the use of acoustical foam pieces [21] or sound-absorbing panels [22]) or light levels (the use of incubator covers [23]) that can cause harm [24].

Accompanied by the incubators, a range of possible devices can be integrated depending on each newborn's need to provide respiratory support, nutritional assistance, medication administration, treatments, and so forth. Moreover, infants in the NICU are under continuous monitoring

of vital signs and cardiac and respiratory activities. Monitoring of these signs enables caregivers to keep track of the infants' condition and initiate immediate and effective care if alarms are triggered.

1.3 Challenges in the NICU

Despite advances in neonatal care and medicine, the NICU continue to face persistent challenges that are often associated with prolonged hospital stays and significantly impact outcomes for the tiniest patients. These challenges include prematurity and low birth weight, infections such as sepsis and Necrotizing enterocolitis, respiratory distress syndromes such as Bronchopulmonary dysplasia, neonatal jaundice, retinopathy of prematurity and neurological disorders, among others [25]. This section will focus on two of these critical challenges: neonatal hyperbilirubinemia and neonatal sepsis.

1.3.1 Neonatal hyperbilirubinemia

Hyperbilirubinemia affects 60% to 80% of newborns [26]. It is characterized by abnormally high levels of bilirubin, a yellow pigment produced during the breakdown of red blood cells, in the blood of newborns. Neonatal hyperbilirubinemia (NHB) can result in neonatal jaundice, a clinical condition marked by the yellow coloration of the skin and the sclera (whites of the eyes) in newborn babies. While most cases of jaundice are common and usually benign—bilirubin itself is considered to act as a powerful antioxidant [27]—the potential neurotoxicity of elevated bilirubin levels necessitates careful monitoring of all newborns and the application of appropriate treatments [28].

Clinical practice guidelines from worldwide organizations, including the American Academy of Pediatrics (AAP) [28–30] in the US and the National Institute for Health and Care Excellence (NICE) [26] in the UK, provide systematic, evidence-based and practical recommendations for the management of neonatal hyperbilirubinemia, aiming to prevent severe hyperbilirubinemia and its associated complications. Many other countries and regions have published guidelines that adapted to their specific national circumstances, such as Australia [31], Canadian [32], China [33], Israel [34], Italy [35], Norway [36], South Africa [37], Spain [38], Switzerland [39], and Turkish [40] etc.

Bilirubin metabolism

The normal metabolism of bilirubin, as shown in Figure 1.3, can be summarized in five main steps [41], including (1) production, (2) uptake by the hepatocyte, (3) conjugation, (4) excretion into bile ducts, and (5) delivery to the intestine.

The predominant production of bilirubin is the breakdown of hemoglobin in senescent or hemolyzed red blood cells. Hemoglobin is degraded by heme oxygenase, resulting in the release

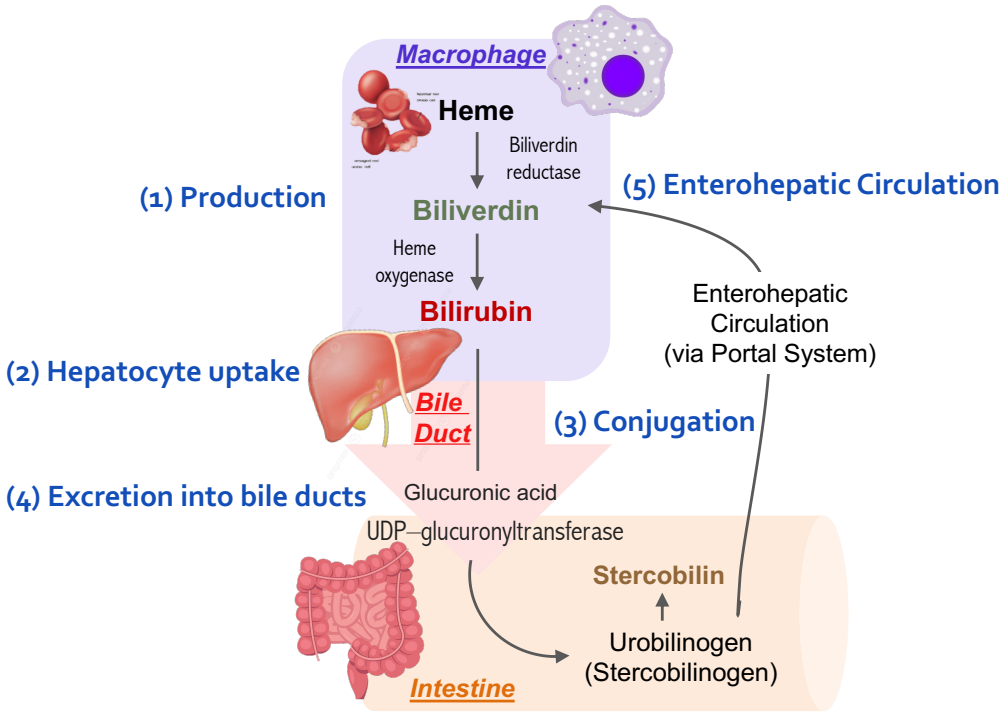

Figure 1.3: Illustration of bilirubin metabolism.

of iron and the formation of carbon monoxide and biliverdin. Biliverdin is further reduced to bilirubin by biliverdin reductase. In this unconjugated (or "indirect") form, bilirubin is water-insoluble and is mostly transported to the liver tightly bound to albumin although some are "free" and hence able to enter the brain. Bilirubin then enters the liver and is uptaken by the hepatocytes and subsequently conjugated with glucuronic acid via the enzyme uridine diphosphate–glucuronyl transferase. In this state, bilirubin is soluble in water and it is called conjugated (or "direct") bilirubin. Then, bilirubin is modified to an excreted bound form which enters the intestinal lumen but can be dissociated by bacteria, thus bilirubin is reabsorbed into the circulation [41, 42].

Causes

Hyperbilirubinemia in term and preterm infants can arise from a variety of factors. Defects or immaturity in any stage of the abovementioned bilirubin metabolism, such as increased production of bilirubin (bilirubin production in neonates is reported to be twice as high as in adults and this is because their blood cells have a shorter lifespan [43]), deficient hepatic uptake, impaired conjugation or increased enterohepatic circulation of bilirubin, can cause high concentrations of bilirubin in the blood, potentially resulting in hyperbilirubinemia [41, 44].

Additional contributing factors include breastfeeding difficulties, blood group incompatibility (commonly Rhesus or ABO incompatibility), other causes of hemolysis, sepsis (infection), genetic

disorders such as deficiency of an enzyme of Glucose-6-phosphate dehydrogenase (G6PD), and metabolic conditions, all of which can exacerbate the condition.

Physiological jaundice

Most infants develop visible jaundice typically during 1-2 days after birth and it usually resolves within 1-2 weeks. This common and generally harmless phenomenon is known as physiological jaundice. It is also proposed that the mild elevation of bilirubin may provide natural antioxidant protection, particularly at a stage in life when other physiological antioxidant defenses are not fully developed [45].

Physiological jaundice is caused by a normal and transient increase in unconjugated bilirubin concentration, which involves the production, excretion, and entero-hepatic circulation of bilirubin, without any underlying pathological cause. During this condition, Total serum bilirubin (TSB) concentrations undergo a rapid rise phase and then a decline to lower levels, eventually mimicking adult values. This pattern is attributed to increased bilirubin production coupled with universally reduced bilirubin conjugation in newborns, and the subsequent equilibrium between bilirubin production and elimination leads to a leveling off or decrease in bilirubin levels over time [30, 43, 46, 47]. However, the peak bilirubin levels, as well as the duration of rising and falling phases, can greatly vary in newborns depending on gestational age, birth weight, and/or the presence of other pathological conditions.

Bilirubin encephalopathy

When bilirubin levels become excessively high in newborns and progress to severe hyperbilirubinemia (TSB levels >20 mg/dL (340 µmol/L)) and extreme hyperbilirubinemia (when TSB >25 or 30 mg/dL (428 or 513 µmol/L)) [48], neonates are particularly vulnerable to bilirubin-induced neurological damage due to their undeveloped blood-brain barrier. This is because elevated unconjugated bilirubin can penetrate the membrane by passive diffusion, increasing neuronal oxidative stress and decreasing neuronal proliferation. The bilirubin neurotoxicity produces damage to the Central nervous system (CNS) and leads to Bilirubin-induced neurologic dysfunction (BIND) through its deposition in selective brain regions including the basal ganglia, hippocampus, and brainstem nuclei [49, 50].

Hyperbilirubinemia affecting the CNS can cause both short-term and long-term neurological dysfunction. Acute bilirubin encephalopathy (ABE) often appears within the first weeks after birth presenting with non-specific clinical symptoms, such as lethargy, hypotonia or hypertonia, poor feeding, fever, a high-pitched cry, seizures, recurrent apnea and desaturations, and even death. If left prolonged and untreated, ABE may ensue and progress to Chronic bilirubin encephalopathy (CBE), also known as kernicterus (yellow staining in the brain). CBE is associated with irreversible and permanent neurological sequelae, including cerebral palsy, movement disorders such

as athetoid palsy and dystonia, auditory neuropathy spectrum disorders such as hearing loss, oculomotor impairments, digestive dysfunction and intellectual disabilities [26, 49, 51, 52].

The exact level of bilirubin likely to cause neurotoxicity in an individual infant varies and depends on the interplay of multiple factors, including acidosis, lower gestational and postnatal age (preterm birth), the rate of serum bilirubin rise, serum albumin concentration, sepsis and significant clinical instability within the previous 24 hours, etc. [26, 28].

Diagnosis

A very intuitive yet subjective method for the preliminary diagnosis of neonatal hyperbilirubinemia is the visual assessment, based on observing the yellowing of the neonatal skin and sclera. However, the serum bilirubin level required to cause jaundice can vary depending on skin tone and the affected body region.

The gold standard for NHB diagnosis involves collecting blood samples, typically through a heel prick or venipuncture, and performing blood tests in the laboratory to measure Total serum bilirubin (TSB) levels. TSB measurements are accurate and reliable, particularly crucial for diagnosing and managing severe hyperbilirubinemia. While this method provides precise results, it can be painful for the infant and cause blood spoliation.

Non-invasive devices such as the Minolta JM-103 and the BiliChek provide Transcutaneous bilirubin (TcB) measurements by emitting a flash of light onto the skin and calculating the amount of bilirubin based on how the light is absorbed or reflected. These devices are becoming common alternatives to TSB testing, offering reasonable estimates for healthy infants [53] and those with significant hyperbilirubinemia [54–60].

Transcutaneous bilirubin meters are quick, painless and cost-effective tools, but their accuracy in estimating TSB at higher bilirubin levels (greater than 15 mg/dL or 256 µmol/L [54, 61, 62]), in preterm and low birth weight infants [63–65], across diverse races and ethnicities [57, 66], and before, during or after phototherapy [64, 66, 67] still requires further investigation.

In clinical practice, as recommended by the NICE guidelines on jaundice in newborns under 28 days [26], TSB measurement should be used for infants with a GA of fewer than 35 weeks or within the first 24 hours of life; TcB measurement is encouraged for infants with a GA of 35 weeks or more and who are older than 24 hours. TSB should also be measured when bilirubin levels meet or exceed relevant treatment thresholds for their age, and for all subsequent measurements. Besides, if TcB measurement indicates a bilirubin level greater than 250 µmol/L (15 mg/dL according to the AAP recommendations [29]) or the infant is under 35 weeks gestation, additional TSB measurements should be conducted to confirm the results.

Treatment

Phototherapy (PT) and Exchange transfusion (ET) are the standard treatments for NHB and can prevent TSB levels from reaching dangerous thresholds.

The use of phototherapy was discovered accidentally in a premature baby ward of Rochford Hospital in Essex, England, by Sister J. Ward, who noticed that the skin of jaundiced infants became bleached when exposed to sunlight. This observation led pediatric pathologist Dr. R. J. Cremer [68] to experiment with blue fluorescent lamps, successfully demonstrating a reduction in bilirubin levels, which marked the birth of phototherapy as a treatment for neonatal jaundice [69]. The blue light used in phototherapy, with wavelengths ranging from 350 nm to 500 nm and peaking near 455 nm, penetrates the skin well and is absorbed maximally by bilirubin, triggering photochemical reactions that convert bilirubin into excretable isomers and breakdown products [70, 71].

The primary goal of phototherapy is to reduce TSB concentrations and prevent further increases that could lead to a need for escalation of care such as exchange transfusion [28]. When TSB concentrations approach dangerously high levels (e.g., diagnosed as ABE) or when intensive phototherapy is ineffective (e.g., hemolysis is occurring), exchange transfusion should then be provided. This procedure involves removing and replacing the infant's blood to rapidly reduce bilirubin levels.

Key aspects of treating hyperbilirubinemia include the timing of treatment initiation, close monitoring and care during treatment, determining when to discontinue therapy, and checking for possible rebound bilirubin levels. The TSB level is consistently used to guide the management of hyperbilirubinemia in all infants.

Initiating treatment

The existing guidelines for managing NHB provide clear thresholds and recommendations to help caregivers decide when to intervene, based on factors such as gestational age, hours since birth, and current TSB concentrations. In addition, the presence of risk factors, including prematurity, sepsis, hemolytic disease, etc., is also taken into account to adjust treatment thresholds accordingly. To illustrate these treatment strategies, we examine two representative guidelines.

The first is the latest version of the AAP guidelines, originally published in 2004 [29], clarified in 2009 [30] and updated in August 2022 [28]. It offers recommendations for phototherapy and exchange transfusion thresholds for neonates born at 35 or more weeks of gestation. Recommended serum bilirubin thresholds for initiating phototherapy and exchange transfusion by gestational age and age in hours after birth are shown in Figure 1.4 and Figure 1.5, respectively. The guidelines also account for hyperbilirubinemia neurotoxicity risk factors.

Another widely practiced consensus-based guideline is from the NICE [26], which offers a more inclusive approach, covering newborns across the entire spectrum: from preterm, full-term

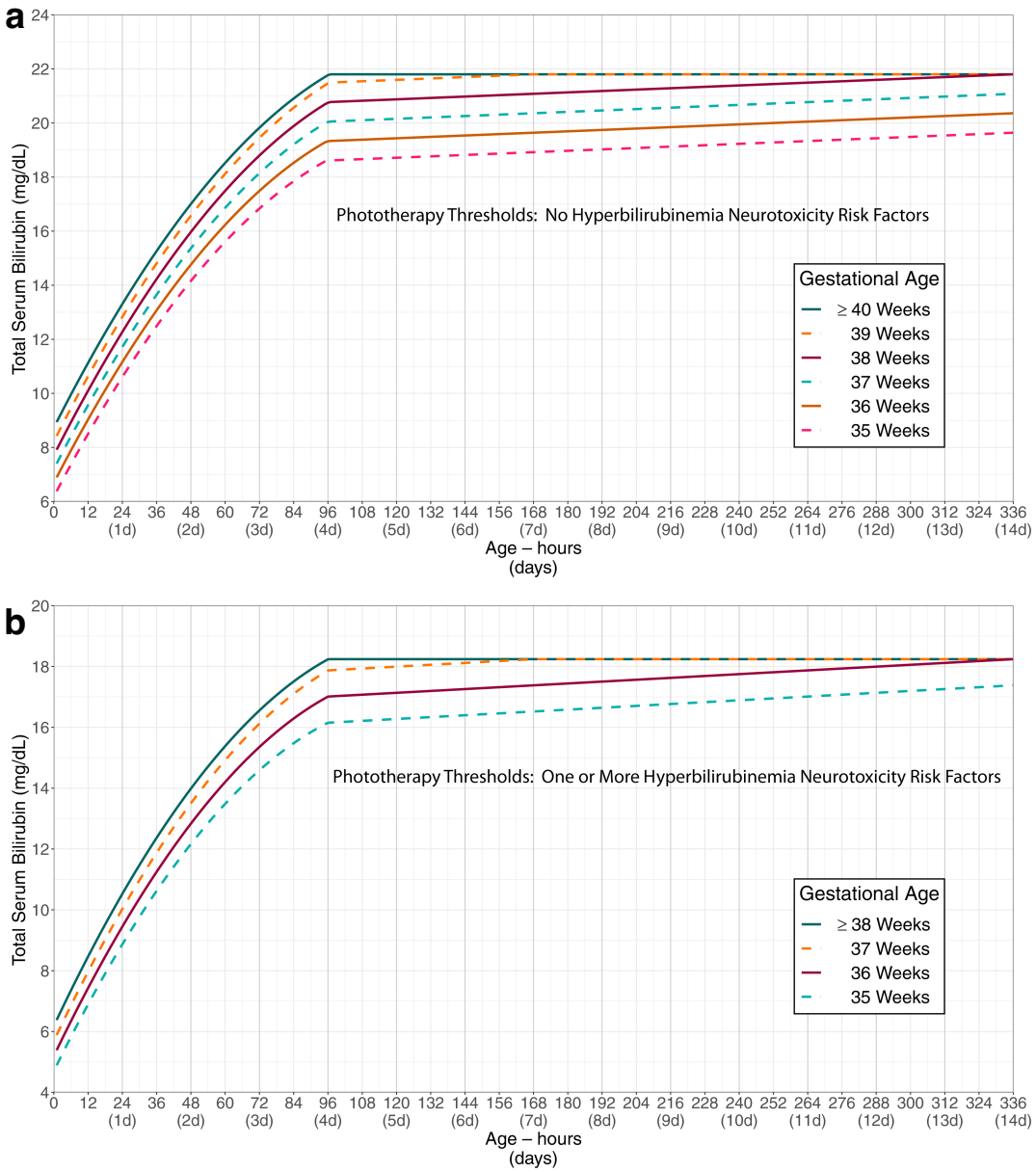

Figure 1.4: Phototherapy thresholds by gestational age and age in hours for infants with (**a**) **no** recognized hyperbilirubinemia neurotoxicity risk factors other than gestational age and (**b**) **any** recognized hyperbilirubinemia neurotoxicity risk factors other than gestational age. Hyperbilirubinemia neurotoxicity risk factors include gestational age <38 weeks; albumin <3.0 g/dL; isoimmune hemolytic disease, glucose-6-phosphate dehydrogenase (G6PD) deficiency, or other hemolytic conditions; sepsis; or any significant clinical instability in the previous 24 hours. Bilirubin 1 mg/dL = 17.1 µmol/L.
(Adapted from [28]. Reproduced with permission from Journal Pediatrics, Vol. 150, Copyright ©2022 by the American Academy of Pediatrics.)

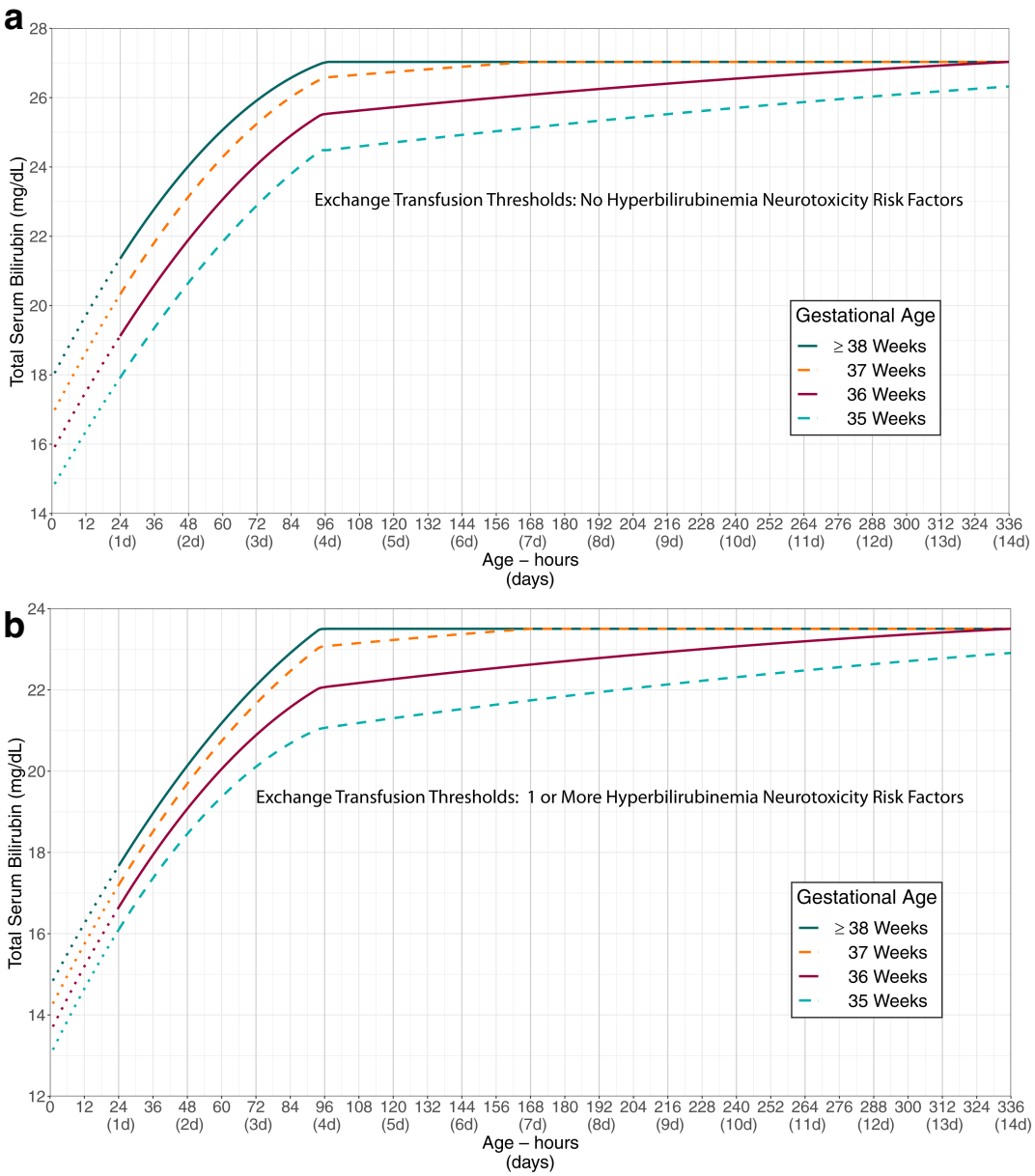

Figure 1.5: Exchange transfusion thresholds by gestational age and age in hours for infants with (**a**) **no** recognized hyperbilirubinemia neurotoxicity risk factors other than gestational age and (**b**) **any** recognized hyperbilirubinemia neurotoxicity risk factors other than gestational age. The stippled lines in (**b**) for the first 24 hours indicate uncertainty because of the wide range of clinical circumstances and responses to intensive phototherapy. Hyperbilirubinemia neurotoxicity risk factors include gestational age <38 weeks; albumin <3.0 g/dL; isoimmune hemolytic disease, glucose-6-phosphate dehydrogenase (G6PD) deficiency, or other hemolytic conditions; sepsis; or any significant clinical instability in the previous 24 hours. Bilirubin 1 mg/dL = 17.1 µmol/L. *(Adapted from [28]. Reproduced with permission from Journal Pediatrics, Vol. 150, Copyright ©2022 by the American Academy of Pediatrics.)*

to post-term births. For infants born at 38 or more weeks of gestation, the threshold for initiating phototherapy is set at 350 µmol/L for those aged 96 hours or more. For infants aged less than 96 hours, the threshold decreases stepwise every 6 hours, starting from 350 µmol/L. The threshold for performing an exchange transfusion is set at 450 µmol/L for infants aged 42 hours or more; while the thresholds in TSB concentrations reduce by 50 µmol/L every 6 hours for those younger than 42 hours. Specific age-after-birth-based thresholds are presented in table form (Table 1.1) and graphic form (Figure 1.6). For preterm infants, thresholds are determined using simple formulas that have been proposed for use in pediatric textbooks for many years, as shown in Table 1.2. It specifies two stages: for infants aged 72 hours or older, GA-sensitive thresholds are calculated using the formula in the second column of table, denoted as $threshold_{PT}^{72h}$ for phototherapy and $threshold_{ET}^{72h}$ for exchange transfusion. For infants younger than 72 hours of life, the thresholds are defined by linear increases, using the formula in the third column of the table. Additionally, to better assist healthcare professionals in determining whether a jaundiced newborn requires treatment, the NICE guidelines provide a series of interactive Excel graphs of TSB versus age in hours, with a separate graph for each gestational age (from 23 weeks to 37 weeks of gestation). An example graph for infants at 32 weeks of gestation is shown in Figure 1.7.

During and discontinuing treatment

There is no unified consensus among existing guidelines regarding the frequency of serum bilirubin monitoring, criteria for discontinuing phototherapy, or recommended intervals for checking for rebound jaundice. But in general, decisions should be guided by the infant's age, the presence of hyperbilirubinemia neurotoxicity risk factors, TSB concentrations, and the TSB trajectory. It is suggested to repeat serum bilirubin measurements 4-6 hours after phototherapy has been applied [26] (or within 12 hours according to AAP [28]) to monitor treatment progress. Once the TSB level is stable or falling, the interval for measurements can be extended to every 6-12 hours.

The decision to discontinue phototherapy involves balancing the desire to minimize both phototherapy exposure and the separation of mothers and infants against the risk of rebound hyperbilirubinemia. A reasonable point to stop phototherapy is when TSB levels fall at least 50 µmol/L (or at least 2 mg/dL suggested by AAP recommendations [28]) below the corresponding treatment threshold. A follow-up TSB measurement is recommended 12-18 hours after stopping phototherapy to check for possible rebound jaundice.

In extreme cases requiring exchange transfusion, continuous multiple phototherapy should be maintained, and a TSB measurement should be taken within 2 hours post-ET, with subsequent management according to the established treatment threshold tables and graphs.

Challenges in hyperbilirubinemia management for preterm infnats

Compared to term and near-term infants, there is less consistent and limited evidence-based guidance for infants born prematurely; the existing guidelines are, by necessity, consensus-based.

Table 1.1: Consensus-based bilirubin thresholds for the management of infants with hyperbilirubinemia born ≥38 weeks of gestation, according to NICE guidelines [26]. *(Adapted from [26].)*

Age (hours)	Bilirubin measurement (micromol/litre)			
0			> 100	> 100
6	> 100	> 112	> 125	> 150
12	> 100	> 125	> 150	> 200
18	> 100	> 137	> 175	> 250
24	> 100	> 150	> 200	> 300
30	> 112	> 162	> 212	> 350
36	> 125	> 175	> 225	> 400
42	> 137	> 187	> 237	> 450
48	> 150	> 200	> 250	> 450
54	> 162	> 212	> 262	> 450
60	> 175	> 225	> 275	> 450
66	> 187	> 237	> 287	> 450
72	> 200	> 250	> 300	> 450
78		> 262	> 312	> 450
84		> 275	> 325	> 450
90		> 287	> 337	> 450
96+		> 300	> 350	> 450
Action	↓	↓	↓	↓
	Repeat bilirubin measurement in 6–12 hours	**Consider phototherapy and repeat bilirubin measurement in 6 hours**	**Start phototherapy**	**Perform an exchange transfusion unless the bilirubin level falls below threshold while the treatment is being prepared**

Table 1.2: Consensus-based bilirubin thresholds for the management of infants with hyperbilirubinemia born <38 weeks of gestation, according to NICE guidelines [26]. *(Adapted from [26].)*

	72 hours or older	Birth to 72 hours of life
Phototherapy thresholds (µmol/L)	(GA in weeks × 10) − 100	$\text{PNA in hours} \times \frac{(\text{thresholds}_{\text{PT}}^{72\text{h}} - 40\ \mu\text{mol/L})}{72\ \text{hours}} + 40\ \mu\text{mol/L}$
Exchange transfusion thresholds (µmol/L)	GA in weeks ×10	$\text{PNA in hours} \times \frac{(\text{thresholds}_{\text{ET}}^{72\text{h}} - 80\ \mu\text{mol/L})}{72\ \text{hours}} + 80\ \mu\text{mol/L}$

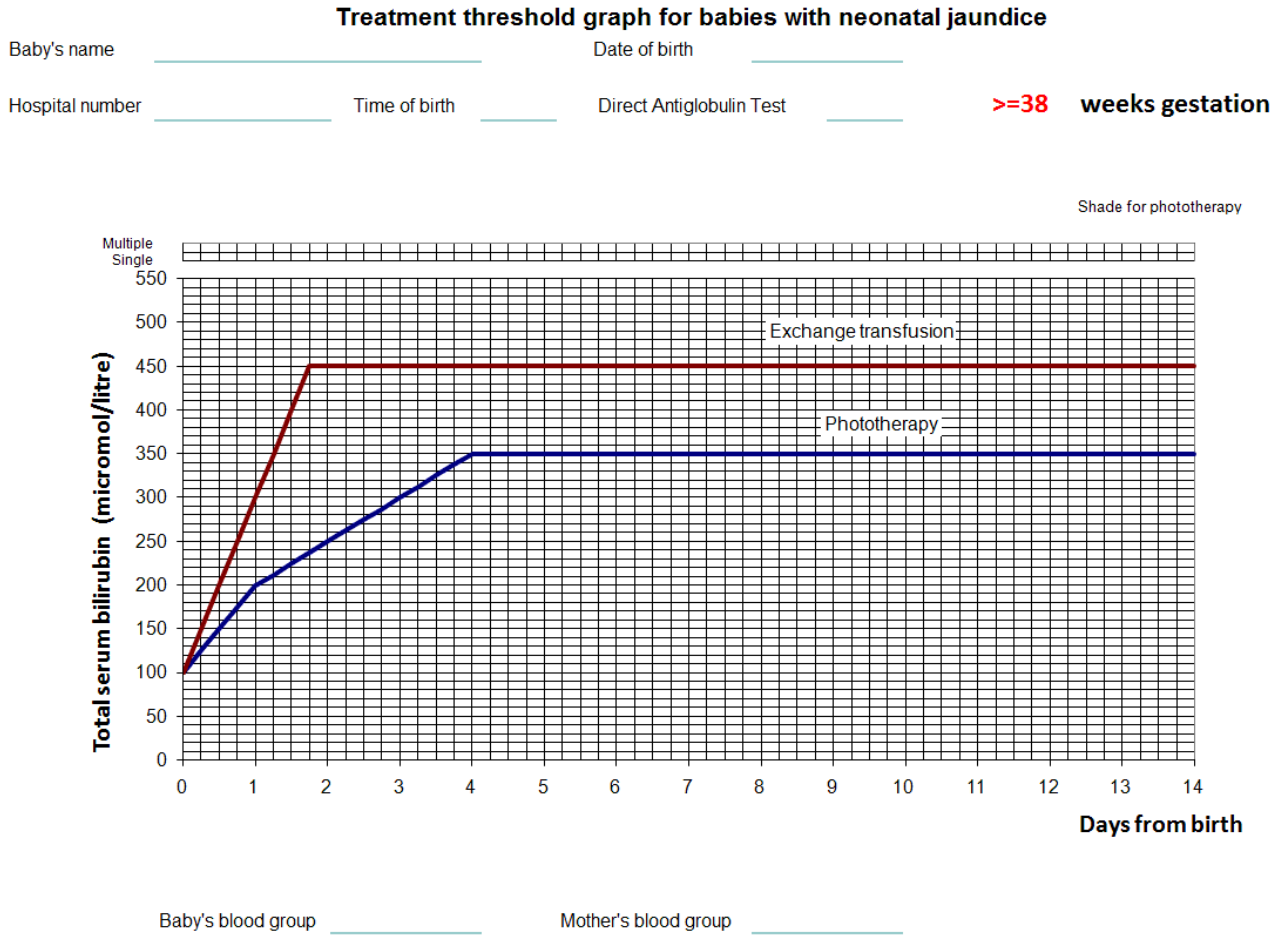

Figure 1.6: Consensus-based bilirubin thresholds for the management of infants with hyperbilirubinemia born ≥38 weeks of gestation, according to NICE guidelines [26]. *(Adapted from [26].)*

Preterm birth (gestational age under 38 weeks) is one of the risk factors leading to significant hyperbilirubinemia [26]. Indeed, hyperbilirubinemia tends to be more prevalent and severe, and its course is more protracted in preterm infants than in infants born term due to their underdeveloped organs and corresponding functions and higher susceptibility to complications [27, 72].

Moreover, the risk of bilirubin-induced neurotoxicity is higher in preterm infants compared to those born term [51, 72, 73]. Several large cohort studies have demonstrated an association between peak TSB levels during the neonatal period (mainly within the first 14 days of life) and adverse neurodevelopmental outcomes at a corrected age of 18 to 30 months (around 2 years). In a cohort study of 1,338 preterm infants (831 babies were available for follow-up) with GA <32 weeks and/or BW <1,500 grams in the Netherlands, Van de Bor et al. [74] found a direct link between TSB levels and abnormal neurological outcomes in these preterm infants when they reached 2 years of old. This association remained evident at 5 years of age in those who had suffered intracranial

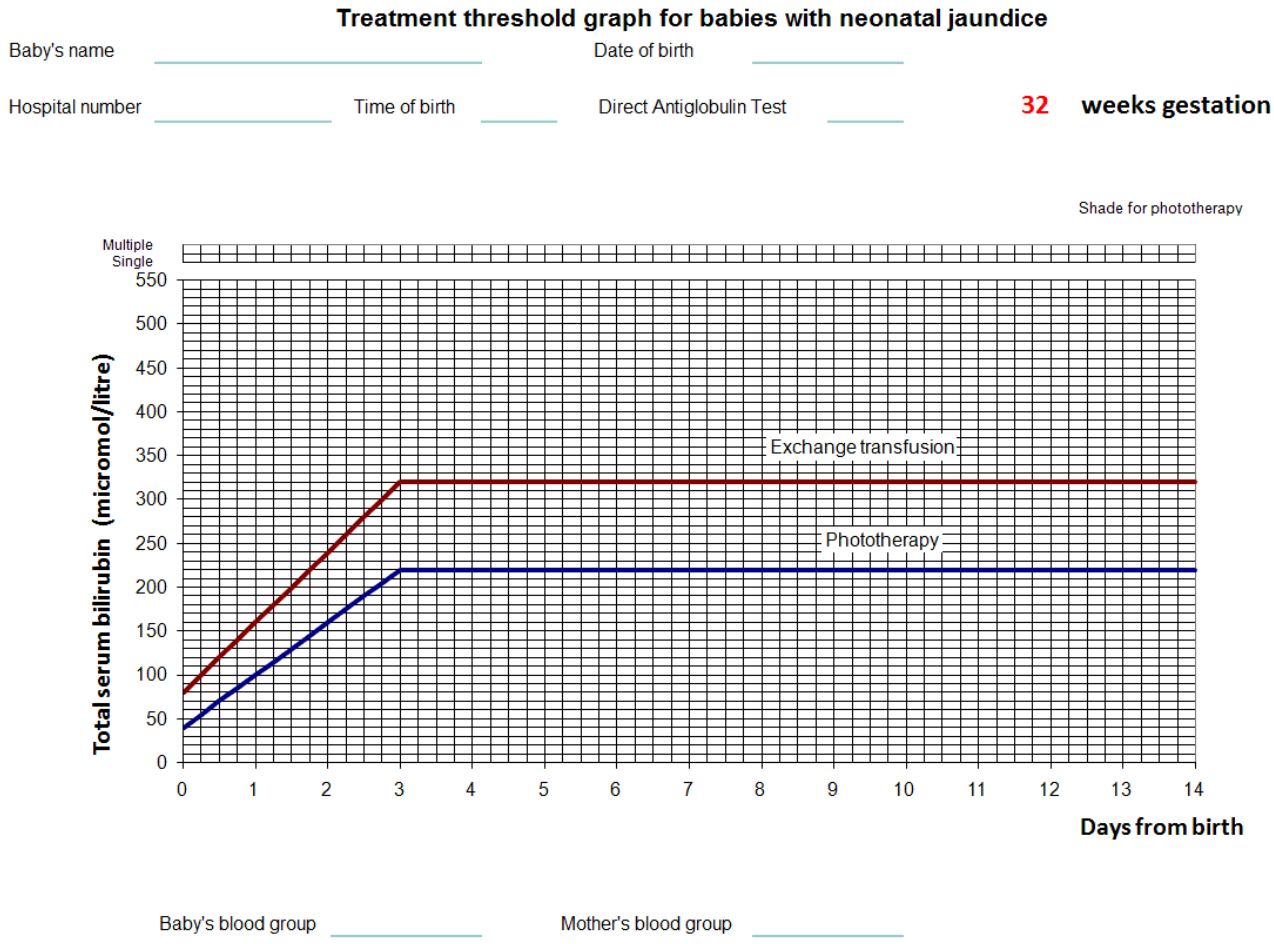

Figure 1.7: Consensus-based bilirubin thresholds for the management of infants with hyperbilirubinemia born at 32 weeks of gestation, according to NICE guidelines [26]. *(Adapted from [26].)*

hemorrhage [74]. Similarly, Oh et al. [75] reported a higher risk of the composite outcome of death or neurodevelopmental impairment (odds ratio: 1.068) and hearing impairment (odds ratio: 1.138) in a cohort of 2,575 infants with follow-up data, drawn from an Amerian cohort of 3,246 extremely low birth weight survivors (401–1,000 grams). A recent multi-center cohort study involving more than 12,000 extremely preterm infants (born between 22 to 28 weeks of gestation, with a median [IQR] birth weight of 920 [750; 1,105] grams) in Canada [76] further supported these findings. Solis-Garcia et al. portrayed normative total peak bilirubin distributions and provided evidence that preterm infants in the highest GA-specific quartile for peak TSB levels had greater odds of neurodevelopmental and hearing impairments. While chronic bilirubin encephalopathy including kernicterus, is currently a rare event in premature neonates, these findings suggest that even subtle elevations in TSB may potentially play a role in abnormal neurodevelopmental outcomes observed in this population [77].

To sum up, improving bilirubin monitoring and management in preterm infants remains a critical issue and an urgent need. This dissertation approaches these challenges from two perspectives.

Firstly, it is essential to examine the evolution of bilirubin levels in preterm infants, not only pre-phototherapy but also post-treatment, to gain a more comprehensive understanding of long-term bilirubin dynamics. Characterizing these dynamics will provide valuable insights into the natural progression and potential complications associated with hyperbilirubinemia in this vulnerable population. By analyzing the bilirubin trends over time, healthcare providers can develop more accurate and individualized treatment protocols.

Secondly, there is a need to validate and implement more non-invasive approaches for managing TSB. Leveraging the advanced monitoring resources and extensive longitudinal data available in NICU, these approaches should be evaluated in the context of massive longitudinal clinical data. This goal is to identify more effective methods for future integration into on-the-edge clinical decision support systems. Such systems, once deployed, may enhance real-time clinical decision-making, enabling timely and effective interventions for preterm infants at risk of severe hyperbilirubinemia and its associated complications.

By focusing on these two areas, we may improve the management of hyperbilirubinemia in preterm infants, ultimately reducing the incidence of bilirubin-induced neurotoxicity and improving overall outcomes.

1.3.2 Neonatal sepsis

Sepsis refers to a dysregulated host response to infection leading to life-threatening organ dysfunction [78]. Specifically, neonatal sepsis is a generalized inflammatory reaction associated with a serious infection in infants under 28 days of age [79, 80]. It is often caused by bacterial, fungal, or viral bloodstream infections. Neonatal sepsis continues to remain a leading cause of neonatal morbidity and mortality, particularly in low- and middle-income countries and regions [81, 82]. Statistically, the incidence varies from 1 to 4 cases per 1,000 live births in high-income countries, but as high as 49-170 cases in low- and middle-income countries with a case fatality rate up to 24% [83]. Survivors of neonatal sepsis are at increased risk for adverse neurodevelopmental outcomes including cerebral palsy, hearing loss, visual impairment and cognitive delays, etc. Moreover, the incidence of sepsis is significantly higher in preterm infants, as well as those with very low birth weight (<1000 grams). Compared to term infants, sepsis in preterm infants is up to 1000-fold more common and is associated with higher rates of mortality and life-long neurodevelopmental disabilities [84, 85].

Categories and causes

Neonatal sepsis is divided into two categories: Early-onset sepsis (EOS) and Late-onset sepsis (LOS). EOS refers to sepsis presenting within the first 72 hours of life (some experts use 7 days), while LOS is defined as sepsis occurring at or after 72 hours of life [85].

EOS is primarily caused by pathogens transmitted vertically from the mother during delivery, such as Group B Streptococcus and Escherichia coli. In addition to prematurity and low birth weight, other key risk factors for EOS include chorioamnionitis, intrapartum fever and premature rupture of membranes [86].

LOS, in contrast, is often acquired postnatally and is linked to nosocomial risk factors, such as contaminated medical devices (e.g., catheters), the hospital environment, or healthcare workers. According to a multi-center survey [87], 21% of Very low birth weight (VLBW) infants who survived beyond 72 hours experienced at least one episode of sepsis. In Extremely low birth weight (ELBW) infants, nearly two-thirds had more than one episode of suspected or culture-proven LOS during hospitalization [88]. Common causative agents include coagulase-negative staphylococci, Staphylococcus aureus, Gram-negative bacteria (such as Klebsiella and Pseudomonas, etc.), and Candida species [89]. In extremely preterm infants (below 32 weeks gestation), the risk of LOS extends beyond the first month of life due to prolonged hospital stays and frequent invasive procedures. Pathophysiologically, the immature immune system is the major contributing factor to increased neonatal susceptibility to sepsis.

The clinical manifestations of neonatal sepsis, whatever EOS or LOS, are often non-specific. Neonates may present with lethargy, temperature instability, poor feeding, apnea, bradycardia, respiratory distress, pneumonia, etc. In particular, the symptoms of LOS tend to be more insidious and non-specific than those of EOS, making early diagnosis rather challenging.

Diagnosis

Early detection is critical, yet no single test can definitively rule in or rule out sepsis. The standard diagnosis of LOS often involves a combination of clinical assessment, laboratory testing and microbiological cultures.

The gold standard for confirming the presence of sepsis is identifying the causative pathogen through blood culture. However, it is invasive, time-consuming, and presents variations in predictive value [90, 91]. Blood exploitation in neonates raises concerns about small blood volume, and obtaining definitive culture results can take days (median [IQR] delay = 21 [13-32] hours [92]). Besides, blood cultures carry the risk of producing misleading results as they are prone to false negatives due to low sensitivity and concurrent antibiotic therapy, as well as false positives due to contamination during culture procedures.

Biomarkers play a valuable role in early diagnosis of sepsis, risk stratification, and guiding

the duration of antibiotic therapy. C-reactive protein (CRP) is an acute-phase reactant that rises in response to inflammation. Although widely used, CRP levels take 6 to 10 hours to rise and 24 to 48 hours to peak after the onset of infection [93], limiting its utility for early diagnosis. Moreover, CRP is non-specific and can be elevated in other inflammatory conditions, making it more useful when used in combination with other tests or for monitoring response to treatment. Procalcitonin (PCT) is more specific to bacterial infections and rises earlier than CRP (within 4–6 hours of infection) and peaks in around 18 to 24 hours [93, 94]. Hence, PCT is considered an early to intermediate-rising biomarker [95]. Other emerging biomarkers, such as Interleukins (IL-6, IL-8) are pro-inflammatory cytokines that rise very early in sepsis, potentially allowing for faster diagnosis [96, 97]. However, both IL-6 and IL-8 have a short half-life and might be affected by non-infection factors, so it is suggested to combine them with later, more specific biomarkers (e.g., CRP) to increase their diagnostic properties [98, 99].

Researchers have also attempted to develop and validate so-called sepsis scores by incorporating different combinations of inflammatory response parameters, laboratory assessments, and physical examination findings, but no single score has proven consistently reliable [82].

Treatment

The treatment of neonatal sepsis, especially LOS, involves prompt initiation of empirical antibiotic therapy to combat suspected bacterial infections. Evidence strongly suggests that prompt initiation of appropriate antibiotic therapy is linked to reduced mortality rates and the risk of severe complications, especially for preterm neonates [100–102]. It has been reported that each hour of delay in initiating antibiotic therapy was associated with an increased mortality of 7.6% [103]. Consideration of early-onset or late-onset presentation and exposures (community versus hospitalized status at the time of symptom onset) affects antimicrobial choice [82]. Common empirical regimens include vancomycin for Gram-positive organisms and gentamicin or third-generation cephalosporins for Gram-negative bacteria. Once pathogen sensitivities are confirmed, definitive antibiotic treatments are tailored accordingly.

In addition to antimicrobial therapy, supportive care plays a critical role, including fluid resuscitation, inotropic support for septic shock, and respiratory support if needed. Monitoring and correcting metabolic imbalances, managing complications like acute kidney injury, and providing antifungal therapy for fungal infections are crucial in more severe cases. Finally, carefully monitoring drug levels and organ function ensures safety and efficacy, and long-term follow-up is needed to assess and address any developmental outcomes.

Currently, several practical and, where possible, evidence-based approaches to the prevention and management of infants with suspected or proven neonatal sepsis have been reported and widely used. While these guidelines vary by country and region, they generally share common principles. Key guidelines include the clinical report for the management of EOS [104], those spe-

cific to neonates born <35 weeks gestation [105] and ≥35 weeks gestation [106] (published by the American Committee on Fetus and Newborn [104]), [107] and the guidelines on neonatal infection aim to reduce delays in recognizing and treating infection and prevent unnecessary use of antibiotics [108] (published by National Institute for Health and Care Excellence (NICE)).

Challenges in LOS management in preterm infants

Despite the reduced burden of neonatal sepsis because of widespread prenatal screening and intrapartum antibiotic administration, missed opportunities for diagnosis and intervention still exist [82]. Prompt and accurate detection and diagnosis are critical for initiating early treatment and improving overall outcomes ultimately.

The most significant challenge in neonatal sepsis management is early detection. The difficulty arises from its non-specific clinical presentation, the limitations of current diagnostic tools, and the vulnerable physiology of neonates. The standard diagnosis method, a positive blood culture, presents several risks. Blood exploitation in neonates raises concerns about small blood volume, and definitive culture results can take days (median [IQR] delay = 21 [13-32] hours [92]). There is also a risk of misleading results from blood cultures, as it is prone to false negatives due to low sensitivity and concurrent antibiotic treatment, as well as false positives due to contamination during the culture process.

Meanwhile, given the rapid progression of infection in this vulnerable population (the course can be fulminant and lead to death within a few hours [109]), clinicians are often compelled to empirically administer antibiotics to infants with risk factors and/or signs of suspected sepsis. Unfortunately, this aggressive and indiscriminate use of antibiotics would expose infants to the risks of antibiotic resistance, unnecessary medication-related adverse outcomes, and also increased healthcare costs [110–113]. Thus, again, the ability to recognize the condition as soon as possible is important, so that targeted and prompt treatment can be started early in the course of infection to prevent further deterioration [101, 102]. At the same time, the development of strategies to reduce antibiotic use and minimize adverse effects, focusing on optimizing the duration of therapy, should be the next frontier for antibiotic stewardship in NICU [82].

All this emphasizes the need for improved diagnostic methods, close clinical monitoring, and timely intervention to reduce morbidity and mortality associated with sepsis. The use of non-invasive and non-culture-based diagnostics and sepsis scores to predict and diagnose septic neonates are areas of active investigation. Moreover, monitoring and assessing long-term outcomes of neonatal sepsis as neonates age remains a notable healthcare challenge [82].

1.4 Clinical Decision Support Systems in NICU

Clinical decision support systems (CDSS), as a subset of clinical informatics, can be defined as "*software that is designed to be a direct aid to clinical decision-making in which the characteristics of an individual patient are matched to a computerized clinical knowledge base, and patient-specific assessments or recommendations are then presented to the clinician and/or the patient for a decision.*" [114]. In such a system, according to Osheroff [115], recipients of the information may include patients, clinicians, and others involved in patient care services; the information provided may include general clinical knowledge and guidance, intelligently processed patient data, or a mix of both; and the format in which the information is delivered may be selected from a rich set of options, including data and order entry aids, filtered data displays, reference messages, alerts, etc. Although definitions vary, the purpose of CDSS is to make it easier for healthcare teams to diagnose, treat, prevent, cure, or mitigate disease and thus improve the overall outcome.

In the following subsections, we introduce the conceptual framework as well as the infrastructure of a CDSS in general. Then we present the features of decision-making in the context of NICU followed by a brief overview of the existing CDSS in NICU. Next, some standards and recommendations for the development and deployment of CDSS are demonstrated. Finally, we conclude with the current challenges with potential benefits and risks related to the CDSS in NICU.

1.4.1 Conceptual framework and infrastructure for CDSS

Most CDSS implementation frameworks incorporate three key components [116–118], as shown in Figure 1.8: *i)* patient data, structured or unstructured data, from one or several sources, *ii)* a knowledge-based (scientific evidence) or non-knowledge-based inference mechanism (e.g., an algorithm, a prediction rule, Bayesian networks, machine learning) and *iii)* a user interaction system.

The patient data includes specific data related to the patient for which the decision will be made. In most cases, the Hospital Information System (HIS) or Electronic health records (EHR) are the main source, where abundant data are integrated through communications with associated systems, including multiple monitoring devices and sensors, radiology, laboratories, pharmacies, scheduling, etc. Clinical data are presented as both structured (e.g., information such as vital signs and diagnosis codes exist within predefined fields) and unstructured data (e.g., information contained within the clinician notes and imaging reports).

Traditional inference mechanisms are knowledge-based and use guidelines and protocols as the basis for decision-making. The domain knowledge and evidence are codified in order to be efficiently and unambiguously executed. A basic strategy is rule-based reasoning, for instance, *if* bilirubin levels surpass a quantitative threshold for a given postnatal age, *then* phototherapy is advised. The knowledge base needs to be updated when new findings are produced. Therefore, the knowledge-based CDSS must be adaptive for adding, deleting and modifying rules. Alternatively,

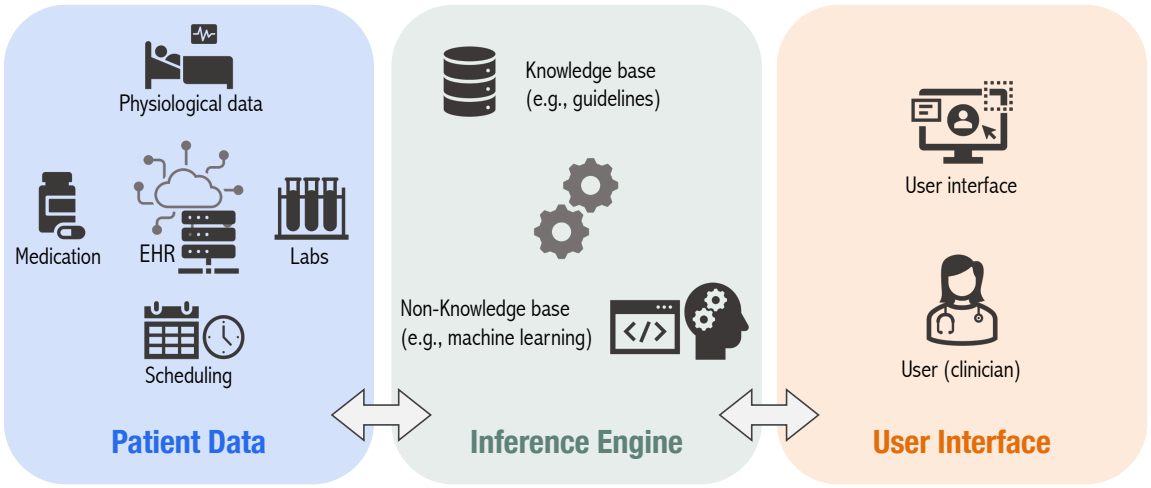

Figure 1.8: Key components of a clinical decision support system. Patient data are from one or several sources and presented in structured or unstructured data formats, they can be gathered through the Electronic health records (EHR) or Hospital Information System (HIS) and directly communicated with multiple devices/sensors. The inference engine draws conclusions through scientific evidence and guidelines (knowledge-based) or complex pattern recognition and analysis (non-knowledge-based). User(s) interact with the input (if any) and output through the user interface.
(Adapted from [117, 118].)

a non-knowledge-based (in the sense that physiological knowledge is not explicitly represented into the model) inference mechanism uses methods such as Machine learning (ML) to generate conclusions from complex patterns that are presented as input to these models. A representative collection is well-known as ML-based or AI-based CDSS. These systems are particularly effective when large amounts of data and multiple variables are used for risk stratification or prediction [118].

The ways in which users interact with CDSS and implement CDSS recommendations differ widely depending on the infrastructure installed in the hospitals. The interface can be a standalone installation limited to a single hospital, connected to the local EHR, or web-based and accessible to all. It is critical to have a clear User interface (UI) that is more than just aesthetically pleasing: it should be simple, clear to understand and avoid a lot of data displays or cognitive overload. The UI must be customized to the end user's task depending on their roles from clinicians and caregivers to support staff (case manager, billing). Overall, workflow integration, flexibility, and adaptability of the UI are required [118].

The availability, adaptability, and scalability of CDSS tools, along with their effective implementation, are directly influenced by the infrastructure and design. Early CDSS are standalone units and not integrated into HIS and EHR, making the process labor-intensive, time-consuming and prone to transcription errors. Then with the development of informatics technologies, the

CDSS are integrated with external systems such as EHR and take the data in EHR as input, making it easier for users and capable of running several CDSS. However, such a merged architecture may pose difficulty in portability since the local instances of EHR are often customized and specific to institutional demands. A more advanced and flexible architecture are service-oriented model as categorized in [118], in which the CDSS are physically separated from the HIS and EHR while "integrated" with them through standardized, service-based interfaces. Here, the handshake between systems is achieved through an interface that uses formal ontologies and vocabularies to encode and represent clinical data and recommendations. This field of medical device interoperability is particularly active both in the research and industrial contexts and its description is out of the scope of this manuscript. The major challenges are usually not technical but related to aspects such as intellectual property protection from different medical device manufacturers or regulatory constraints.

Overall, it is clear that architecture, or infrastructure, is essential to the development of a scalable, EHR-independent platform that can operate in a variety of settings. Platform independence can be achieved by leveraging the capabilities of data representation, exchange, and storage standards, including terminology standards (SNOMED, LOINC and RxNorm) [119], information model standards (HL7 [120]) and standards for creating, representing and implementing CDSS interventions [121].

1.4.2 Decison-making in NICU

The processes and medical rules followed within NICU are complicated and more complex and frequent data records are produced due to the dynamic, heterogeneous, and real-time environment of the NICU [122]. In fact, making decisions in the NICU is a difficult task and it revolves around addressing a myriad of crucial aspects. It ranges from determining medical interventions and treatment plans to optimizing monitoring parameters and, indeed, devising overall high-quality and patient-specific care strategies. This multifaceted decision-making process not only affects the quality of newborn care delivery but may also impose psychological burdens on parents and exert financial costs for the broader healthcare system. Most deployed decision support systems incorporating threshold-based alert mechanisms appear far insufficient to encompass the nuanced complexities of neonatal health. These basic alarms often lead to false positives and fail to detect subtle yet critical changes that may indicate a deterioration in an infant's condition.

Achieving the delicate balance demanded to meet the diverse needs of newborns requires clinical expertise and a sharp grasp of real-time data. Gratefully, the last decades have witnessed a surge in technological advancement in artificial intelligence and an era of big data, especially in the medical and healthcare sectors. At the intersection of healthcare and technology, CDSS, especially AI-based CDSS emerge as powerful tools in the management of complex medical conditions. They are designed to assist healthcare providers in making informed decisions and ensuring the best

possible outcomes for patients by improving the quality of patient care, lowering medical errors, and enhancing overall health delivery.

On the other hand, preterm infants in the NICU share the characteristics of "big data", i.e., "high volume, high velocity and high variety" [123], as they generate extensive and high-dimensional data through continuous monitoring. In the context of NICU, the "volume" refers to the massive size of the dataset, and "velocity" refers to the rapid growth of data. "Variaty" refers to the heterogeneous data, which partly comes from continuous monitoring of physiological signals and patients' vital signs, such as heart rate, respiration rate and oxygen saturation [124]. The clinical information collected in Electronic health records (EHR) is another major data source, which covers patient admission information, nutrition status, medication management, laboratory test results, and additional tests like Electroencephalogram (EEG) and magnetic resonance imaging (MRI). Moreover, NICU infants have medical records that start at birth and thus provide a complete and high-fidelity picture of their clinical state [125]. Such a wealth of healthcare data offers an opportunity for Artificial Intelligence and Machine learning applications to flourish in this field.

As a consequence, these conditions hit it off perfectly and the AI-based CDSS hold significant promise within the context of NICU. We present a brief review of the studies of CDSS in the NICU in the following section.

1.4.3 CDSS in NICU: a brief review

A variety of CDSS applications have been developed and to assist in in neonatal care with different objectives, such as hyperbilirubinemia management, medication management, optimizing nutrition, risk estimators for morbidity, mortality, and sepsis [118]. Broadly, these tools leverage data, evidence-based information and statistical analysis or machine learning algorithms to augment the decision-making capabilities of medical professionals in routine care, risk stratification, outcome prediction and treatment plan optimization and personalization toward the unique needs of each neonate.

Neonatal pharmacokinetic and pharmacodynamic present unique challenges due to large variability in their conditions and metabolism functions that necessitate precise medication administration. Simulations, based on 564 gentamicin concentrations among 339 patients with a mean gestational age of 35 weeks (52% preterm births), have shown the utility of a Bayesian CDSS [126] in personalizing gentamicin therapies concerning regimen initiation and response to measured drug concentrations for newborns. Another model-based dosing approach named NeoVanco [127] has been developed and assessed to individualize empiric vancomycin dosing that improves the achievement of target exposure levels in neonates.

Many studies have also reported the effectiveness of ML-based CDSS on antibiotic stewardship [128]. One case is the development of a CDSS based on a supervised learning module to assist

antimicrobial stewardship pharmacists in identifying and reporting potentially inappropriate prescriptions by extracting expert rules for multiple types of antimicrobial alerts [129, 130]. These rules can extend the knowledge base of the baseline system and thus enhance the overall antimicrobial management programs. A recent study [131] developed and evaluated the ability of a ML-CDSS based on categorical boosting to assist clinicians in selecting optimal β-lactam antibiotic doses. The prediction accuracy of all five drugs was >80.0% in the real-world validation.

In hyperbilirubinemia management, many CDSS tools based on guidelines or machine learning techniques have been employed to improve the overall outcomes of jaundiced infants. A widely used web-based BiliTool [132] and its EHR-integrated version [133] follow AAP guidelines. Another web-based Premie BiliRecs [134, 135] was reported to improve adherence to phototherapy guidelines in preterm infants without increased adverse events. Besides, an influence diagram-based CDSS [136] was developed to assist clinicians with making decisions in admission and treatment for jaundice neonates. The use of this CDSS has included a profound change in daily medical practice and avoided aggressive therapies.

Neonatal sepsis including early-onset (EOS) and late-onset sepsis (LOS) is another formidable challenge in NICU. Rich efforts have been made to augment decision-making regarding the risk stratification of EOS [137, 138], early detection of LOS that based on machine learning algorithms using EHR data [139, 140] and HRV [141], etc. A famous monitoring system HeRO [142], based on the study by Griffin et al. [143], is proposed to evaluate the risk of sepsis in real-time by receiving and processing ECG data monitored in existing NICU bedside monitors. It is reported to successfully decrease the mortality rate of the HeRO group by 22% than the controls [144].

In fields of diagnosis, risk assessment and outcome prediction, the CDSS especially AI- or ML-based approaches have been diversely used.

An AI-assisted detection of Patent ductus arteriosus (PDA) from neonatal phonocardiogram was proposed in [145]. It pre-processes and segments the heart sounds and extracts features to be input into a boosted decision tree classifier to estimate the probability of PDA. The evaluation of the proposed system was conducted on a large clinical dataset of heart sounds from 265 term and late-preterm newborns recorded within the first six days of life, and an AUROC of 78% was obtained. The performance for PDA detection compares favorably with the level of accuracy achieved by an experienced neonatologist.

Extubation failure is another ongoing problem in NICU. Mikhno et al. proposed a prediction algorithm to differentiate between patients with successful extubation and those with failed extubation [146]. Their algorithm was a logistic regression model using six features as input, and the performance had an AUROC of 87.1% with a sensitivity of 70.1% and a specificity of 90%. The use of light and routinely recorded variables allows for the possibility of developing a real-time CDSS.

CDSS were also employed in neonatal encephalopathy detection and severity grading. Temko

et al. developed an automated neonatal seizure detection system based on Support vector machine (SVM) classifiers receiving multi-channel EEG as input [147]. This system was validated on both large clinical datasets [148] and randomized datasets [147], achieving a mean Area under the receiver operating characteristic curve (AUROC) of over 95.4%. Raurale et al. [149] proposed a system combining a quadratic time-frequency distribution with a convolutional neural network CNN that classifies four EEG grades of Hypoxic-ischaemic encephalopathy injury (HIE). The system achieves an accuracy of 88.9% on the development dataset and 69.5% on a large unseen test dataset.

There are a range of CDSS targeting mortality risk assessment. For instance, novel real-time models using Decision Tree (DT) and ANN were developed [150]. The models met the criteria for clinically useful results (>60% sensitivity, >90% specificity) for individual patients, opening the potential for a prototype of the CDSS framework to be implemented with the ability to generate intelligent alerts and warnings from medical events as they occur, such as when the risk estimation for a patient changes significantly.

In summary, CDSS in the NICU are advanced tools designed to assist healthcare professionals in making more informed and timely decisions regarding the care of critically ill newborns. These systems integrate patient data, such as vital signs, lab results, electrophysiological signals, and medical history, with evidence-based guidelines and machine learning algorithms to provide real-time recommendations and alerts. These tools aim to enhance diagnostic accuracy, optimize treatment plans, and reduce the risk of human error by offering insights that may not be immediately apparent to clinicians. Ultimately, these systems support the delivery of personalized and precise neonatal care, improving outcomes for vulnerable infants.

1.4.4 Standards and priorities in AI-based CDSS deployment

Despite the potential of CDSS to improve patient care and patient safety significantly, their successful deployment has been limited. At all stages of the development and deployment of a new CDSS, several factors need to be considered to increase the likelihood of its integration into healthcare services and thus make it a successful deployment.

Firstly, regarding the model development, it is best performed by an interdisciplinary team of stakeholders, including clinicians and other potential end users, data scientists, clinical informaticians, and implementation scientists [124]. The minimum information about clinical artificial intelligence modeling criterion [151] suggests considerations specific to AI.

The second point refers to the standards of the data entry and decision algorithms [152]. Some CDSS that are not fully computerized require users to manually enter patient data, which is time-consuming, labor-intensive and prone to transcription errors or even disruptive to the delivery of patient care. Manual data entry can be minimized by integrating CDSS with HIS and EHR. Another

major point is keeping the CDSS decision-making algorithms up-to-date as patient management changes from time to time [152]. The system can be designed to retrain itself automatically and periodically, or it can be designed as a plug-in to facilitate updates.

Thirdly, the human-computer interaction, including data acquisition and the manner of information requests from the system, should be clear, simple, and easily accessible while secure [152]. The UI should be user-friendly, intuitive and offer easy access to information. To respond to emergency situations that often arise in the NICU, the CDSS should be designed to use the least amount of clinician time possible; this includes time to log in to the system and time to acquire the information desired [153]. In addition, it would be more convenient for clinicians if information could be obtained from mobile CDSS or CDSS with many terminals, rather than from a single terminal that may be located a long distance away [153].

Lastly, the CDSS must fit into the clinician's workflow and provide them with useful information. The format and type of system output depend on the clinician's needs. Each clinician has different work habits and therefore may require different functionality. This makes the development of an effective CDSS more complex, but it is important for a successful deployment [152]. Achieving this involves the necessity of close cooperation with the users at every step in the deployment of the CDSS. Besides, Desirable attributes of CDSS include smart information and smart alerts, which involve a subtle balance between too many and too few. It is important that the CDSS be able to anticipate the need for information and deliver it in real-time without clinicians needing to explicitly ask for it [154].

A recent editorial [155] identified important priorities that must be incorporated into CDSS in order to be accepted and integrated into the routine clinical workflow:

- Black box are unacceptable.
  A great portion of AI algorithms are "black box" which decisions are made by algorithms that lack interpretability, such as deep learning networks. In the clinical context, transparency and clinician trust in the models are paramount. Thus, the reasoning behind the AI-based CDSS should be transparent in order for the clinician to comprehend the rationale behind the decision.
- Time is scarce resource.
  CDSS should be efficient in terms of time requirements and must blend seamlessly into the workflow of busy clinical environments.
- Complexity and lack of usability thwart use.
  CDSS tools should be intuitively constructed and simple to use so that no major training is required for their use, which emphasizes again the importance of good human-computer interaction design.
- Relevance and insight are essential.
  A CDSS should reflect an understanding of the pertinent domain and answer clinically rel-

evant questions.

- Delivery of knowledge and information must be respectful.
  Advice and suggested decisions should be provided in a way that recognizes the expertise of the user to augment—but not replace—decision-making. In other words, CDSS should be designed with perceived usefulness [156], informing and assisting but not to replace clinicians.
- Scientific foundation must be strong.
  A CDSS should be built on the basis of rigorous, peer-reviewed scientific evidence to establish its safety, validity, reproducibility, usability and reliability.

1.4.5 Challenges of AI-based CDSS in NICU

While AI-based CDSS hold great promise for enhancing care in the NICU, their implementation is not without challenges. Major difficulties concern the development, implementation and credibility.

First, one essential challenge lies in the relative lack of large, high-resolution, high-quality datasets, which are required in the development and generalization of AI models. Though massive data are generated every second in NICU, neonatal can be sparse or inconsistent resulting in low data availability and poor data quality. Imbalanced datasets also pose difficulties in ML approaches and appropriate balancing strategies should be well applied to the raw datasets before being used as input to learning models. When using smaller datasets, sophisticated and advanced AI methods may offer no advantage or merely a slight advantage over simpler and classical Machine learning or shallow neural networks.

Another limitation lies in the lack of the integration of real-time CDSS into the complex and dynamic environment of the NICU, where patient conditions can change rapidly, requiring real-time updates and highly accurate predictions. Applications and frameworks for real-time physiological data analyses are already gaining ground [118]. A few examples include Artemis [157], Baby Steps [158], Etiometry [159, 160] and iNICU [161, 162]. However, it is important to examine and deploy new temporal data mining approaches and system architectures. Significant adaptations of developed models may need to be incorporated into existing neonatal care protocols and systems, including training healthcare professionals to interpret and act upon model outputs effectively. Future work should explore strategies to streamline the deployment of these models in clinical settings, such as developing more efficient algorithms and ensuring interoperability with current healthcare technologies. Meanwhile, robust infrastructure, such as multi-agent systems, services, and sensors, should be established to provide integrated real-time solutions for NICU [122].

Additionally, there are concerns about the explainability and interpretability of AI-driven recommendations, as healthcare providers may find it difficult to trust or understand the rationale behind the decisions made by these systems. The results derive from the “black box”, which flies

in the face of traditional clinical decision-making that relies on reasoning and judgments based on knowledge and experiences. In order to improve credibility and acceptance, the methods need to be easy to readily explainable, interpretable and transparent [163]. It is suggested to consider the simpler, more applicable, and transparent model over the more complicated one when models perform similarly [124]. Moreover, more analysis such as interpretability analysis and sensitivity analysis should be incorporated to characterize and improve the decision explainability.

Focusing on the areas of concern of this dissertation, machine learning has the potential to transform preterm infant monitoring by analyzing large datasets from Electronic health records (EHR), vital sign data, and other sources of information. However, this field is still in its infancy, facing several challenges [164–166]:

- Data scarcity: Preterm infants are a vulnerable population with limited data availability compared to adult patients or healthy term-born babies.
- Complexity of physiological signals: Vital signs from preterm infants can be noisy and unreliable due to their immature physiology and the need for life-sustaining interventions.
- Limited understanding of disease mechanisms: The pathophysiology of, for example, LOS in preterm infants is not yet fully understood, making it challenging to develop effective machine learning models.
- High stakes: Any errors or misclassifications can have severe consequences, emphasizing the importance of rigorous model development and validation.
- Limited data quality: Develop standardized datasets with high-quality vital sign data and EHR.
- Limited model interpretability: Provide transparent explanations of the decision-making process behind machine learning models.
- Lack of model validation in diverse populations: Test models on different preterm infant cohorts, considering factors like gestational age, birth weight, and comorbidities.

Particularly, this dissertation is interested in a set of major methodological challenges inherent to working with Machine learning (ML) models on NICU neonatal monitoring data, which are longitudinal, continuous, time-dependent, and often noisy. Key challenges often overseen in this context include:

- Temporal dependencies & time-varying covariates:
  Longitudinal data often exhibits temporal dependencies between observations, and covariates (predictor variables) can change over time, making it challenging for traditional feature engineering and machine learning algorithms, that assume independence.
- Missing values:
  Longitudinal data often contain missing values due to measurement errors or non-response rates, which can lead to biased estimates if not handled properly.

- Temporal granularity & non-stationarity:
  Data may be collected at varying time scales and exhibit time-series properties like seasonality or trends, while patterns and distributions can shift over time, requiring adaptive modeling.
- Model complexity & scalability:
  The complexity of longitudinal data may increase model complexity, raising the risk of overfitting and making it computationally expensive, thus demanding efficient regularization techniques and scalable models.
- Model interpretability:
  Understanding how models use temporal information and how this affects predictions is critical to ensuring they are meaningful and actionable.

1.5 Conclusion

Preterm infants are a vulnerable population due to their overall immaturity, and they suffer various high-risk and life-threatening conditions during their first days and weeks of life. The NICU play a critical role in ensuring their survival and improving long-term outcomes. Among the major challenges faced by preterm infants, neonatal hyperbilirubinemia and neonatal sepsis were highlighted as significant concerns due to their high prevalence and potential for adverse complications and outcomes. On the other hand, as advances in neonatal care continue to evolve and massive amounts of healthcare monitoring data are generated, Clinical decision support systems (CDSS) in NICU are emerging as valuable tools for enhancing clinical decision-making and improving outcomes. These systems offer the potential to assist healthcare providers in coping with clinical challenges by integrating data and providing evidence-based recommendations. However, a number of major challenges persist in this field. By acknowledging these challenges and limitations, we can work towards developing more effective machine learning solutions for preterm infant monitoring, ultimately improving patient outcomes and reducing morbidity and mortality rates. This framework sets the stage for exploring solutions to these challenges in subsequent chapters.

BIBLIOGRAPHY

[1] American College of Obstetricians and Gynecologists et al., "Definition of term pregnancy. committee opinion no. 579," Obstet Gynecol, vol. 122, no. 5, pp. 1139–1140, 2013.

[2] Definitions, WHO Recommended, "Terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths," Acta Obstet Gynecol Scand, vol. 56, no. 3, pp. 247–53, 1977.

[3] World Health Organization, Born too soon: decade of action on preterm birth. World Health Organization, 2023.

[4] E. O. Ohuma, A.-B. Moller, E. Bradley, S. Chakwera, L. Hussain-Alkhateeb, A. Lewin, Y. B. Okwaraji, W. R. Mahanani, E. W. Johansson, T. Lavin et al., "National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis," The Lancet, vol. 402, no. 10409, pp. 1261–1271, 2023.

[5] H. Cinelli, N. Lelong, and C. Le Ray, "Enquête nationale périnatale. rapport 2021," 2022.

[6] A. Ego, C. Prunet, E. Lebreton, B. Blondel, M. Kaminski, F. Goffinet, and J. Zeitlin, "Courbes de croissance in utero ajustées et non ajustées adaptées à la population française. i–méthodes de construction," Journal De Gynécologie Obstétrique Et Biologie De La Reproduction, vol. 45, no. 2, pp. 155–164, 2016.

[7] World Health Organization, "Preterm birth," https://www.who.int/en/news-room/fact-sheets/detail/preterm-birth, 2010, Accessed: 2024-01-02.

[8] H. A. Frey and M. A. Klebanoff, "The epidemiology, etiology, and costs of preterm birth," Seminars in Fetal and Neonatal Medicine, vol. 21, no. 2, pp. 68–73, 2016, prediction and prevention of preterm birth and its sequelae.

[9] W. A. Engle et al., "Age terminology during the perinatal period." Pediatrics, vol. 114, no. 5, pp. 1362–1364, 2004.

[10] J. Perin, A. Mulick, D. Yeung, F. Villavicencio, G. Lopez, K. L. Strong, D. Prieto-Merino, S. Cousens, R. E. Black, and L. Liu, "Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the sustainable development goals," The Lancet Child & Adolescent Health, vol. 6, no. 2, pp. 106–115, 2022.

[11] R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, "Epidemiology and causes of preterm birth," The lancet, vol. 371, no. 9606, pp. 75–84, 2008.

[12] UNICEF DATA, "Levels and trends in child mortality: Report 2022," UN Inter-agency Group for Child Mortality Estimation, Tech. Rep., 2023.

[13] World Health Organization, "Standards for improving the quality of care for small and sick newborns in health facilities.[Internet]," World Health Organization, 2020.

[14] Committee on Fetus and Newborn, W. D. Barfield, L.-A. Papile, J. E. Baley, W. Benitz, J. Cummings, W. A. Carlo, P. Kumar, R. A. Polin, R. C. Tan et al., "Levels of neonatal care," Pediatrics, vol. 130, no. 3, pp. 587–597, 2012.

[15] R. E. Behrman and A. S. Butler, Preterm birth: causes, consequences, and prevention. Washington, DC: The National Academies Press (US), 2007.

[16] J. P. Baker, "The incubator and the medical discovery of the premature infant," Journal of Perinatology, vol. 20, no. 5, pp. 321–328, 2000.

[17] V. C. Smith, K. Love, and E. Goyer, "NICU discharge preparation and transition planning: guidelines and recommendations," Journal of Perinatology, vol. 42, no. Suppl 1, pp. 7–21, 2022.

[18] M. Cox, "When can my baby go home? Milestones NICU babies need to meet before being discharged | UofL Health," https://uoflhealth.org/articles/when-can-my-baby-go-home-milestones-nicu-babies-need-to-meet-before-being-discharged/, 2019, Accessed: 2024-02-29.

[19] A. L. Jefferies and Canadian Paediatric Society and Fetus and Newborn Committee, "Going home: facilitating discharge of the preterm infant," Paediatrics & child health, vol. 19, no. 1, pp. 31–36, 2014.

[20] N. A. Bokulich, D. A. Mills, and M. A. Underwood, "Surface microbes in the neonatal intensive care unit: changes with routine cleaning and over time," Journal of clinical microbiology, vol. 51, no. 8, pp. 2617–2624, 2013.

[21] A. N. Johnson and J. Hayes, "Practice applications of research. neonatal response to control of noise inside the incubator." Pediatric Nursing, vol. 27, no. 6, 2001.

[22] C. V. Bellieni, G. Buonocore, I. Pinto, N. Stacchini, D. M. Cordelli, and F. Bagnoli, "Use of sound-absorbing panel to reduce noisy incubator reverberating effects," Neonatology, vol. 84, no. 4, pp. 293–296, 2003.

[23] L. Hellström-Westas, M. Inghammar, K. Isaksson, I. Rosén, and K. Stjernqvist, "Short-term effects of incubator covers on quiet sleep in stable premature infants," Acta Paediatrica, vol. 90, no. 9, pp. 1004–1008, 2001.

[24] R. Antonucci, A. Porcella, and V. Fanos, "The infant incubator in the neonatal intensive care unit: unresolved issues and future developments," Journal of perinatal medicine, vol. 37, no. 6, pp. 587–598, 2009.

[25] M. Fu, W. Song, G. Yu, Y. Yu, and Q. Yang, "Risk factors for length of NICU stay of newborns: A systematic review," Frontiers in pediatrics, vol. 11, p. 1121406, 2023.

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

[27] R. Stocker, Y. Yamamoto, A. F. McDonagh, A. N. Glazer, and B. N. Ames, "Bilirubin is an antioxidant of possible physiological importance," Science, vol. 235, no. 4792, pp. 1043–1046, 1987.

[28] A. R. Kemper, T. B. Newman, J. L. Slaughter, M. J. Maisels, J. F. Watchko, S. M. Downs, R. W. Grout, D. G. Bundy, A. R. Stark, D. L. Bogen et al., "Clinical practice guideline revision: management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation," Pediatrics, vol. 150, no. 3, 2022.

[29] American Academy of Pediatrics Subcommittee on Hyperbilirubinemia, "Management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation," Pediatrics, vol. 114, no. 1, pp. 297–316, 2004.

[30] M. J. Maisels, V. K. Bhutani, D. Bogen, T. B. Newman, A. R. Stark, and J. F. Watchko, "Hyperbilirubinemia in the newborn infant ≥35 weeks' gestation: an update with clarifications," Pediatrics, vol. 124, no. 4, pp. 1193–1198, 2009.

[31] Queensland Clinical Guidelines, "Queensland clinical guidelines: Neonatal jaundice. Guideline No. MN22.7-V9-R27." http://www.health.qld.gov.au/qcg, December 2022, Accessed: 2024-07-26.

[32] K. J. Barrington, K. Sankaran, C. P. Society, Fetus, and N. Committee, "Guidelines for detection, management and prevention of hyperbilirubinemia in term and late preterm newborn infants," Paediatrics & Child Health, vol. 12, no. suppl_B, pp. 1B–12B, 2007.

[33] Neonatology, Subspecialty Group, Editorial Board of Chinese Journal of Pediatrics and Chinese Medical Association et al., "Experts consensus on principles for diagnosis and treatment of neonatal jaundice," Zhonghua er ke za zhi= Chinese journal of pediatrics, vol. 48, no. 9, pp. 685–686, 2010.

[34] M. Kaplan, P. Merlob, and R. Regev, "Israel guidelines for the management of neonatal hyperbilirubinemia and prevention of kernicterus," Journal of Perinatology, vol. 28, no. 6, pp. 389–397, 2008.

[35] C. Romagnoli, G. Barone, S. Pratesi, F. Raimondi, L. Capasso, E. Zecca, and C. Dani, "Italian guidelines for management and treatment of hyperbilirubinaemia of newborn infants ≥35 weeks' gestational age," Italian journal of pediatrics, vol. 40, pp. 1–8, 2014.

[36] D. Bratlid, B. Nakstad, and T. Hansen, "National guidelines for treatment of jaundice in the newborn: guidelines for treatment of neonatal jaundice," Acta Paediatrica, vol. 100, no. 4, pp. 499–505, 2011.

[37] A. R. Horn, G. F. Kirsten, S. M. Kroon, P. A. Henning, G. Moller, C. Pieper, M. Adhikari, P. Cooper, B. Hoek, S. Delport, M. Nazo, and B. Mawela, "Phototherapy and exchange transfusion for neonatal hyperbilirubinaemia: neonatal academic hospitals' consensus guidelines for South African hospitals and primary care facilities," South African Medical Journal, vol. 96, no. 9, pp. 819–824, 2006.

[38] M. D. S.-R. Sánchez-Gabriel, J. L. L. Castellanos, I. B. Fernández, A. P. Muñuzuri, S. R. Gracia, C. W. R. Campillo, E. S. López, M. S. Luna, C. de Estándares de la Sociedad Española de Neonatología et al., "Guidelines for prevention, detection and management of hyperbilirubinaemia in newborns of 35 or more weeks of gestation," Anales de Pediatría (English Edition), vol. 87, no. 5, pp. 294–e1, 2017.

[39] R. Arlettaz, A. Blumberg, L. Buetti, H. Fahnenstich, D. Mieth, and M. Roth-Kleiner, "Assessment and treatment of jaundice newborn infants $35^{0/7}$ or more weeks of gestation," Swiss Soc Neonatology, vol. 1, no. 4, 2007.

[40] A. Çoban, M. K. Türkmen, and T. Gürsoy, "Turkish neonatal society guideline to the approach, follow-up, and treatment of neonatal jaundice," Turkish Archives of Pediatrics/Türk Pediatri Arşivi, vol. 53, no. Suppl 1, p. S172, 2018.

[41] A. Nakeeb, H. Pitt, and J. Toouli, "Biliary tract pathophysiology," in Surgery of the Liver, Biliary Tract and Pancreas: Volumes 1-2, Fourth Edition. Elsevier, 2006, pp. 79–97.

[42] P. A. Dennery, D. S. Seidman, and D. K. Stevenson, "Neonatal hyperbilirubinemia," New England Journal of Medicine, vol. 344, no. 8, pp. 581–590, 2001.

[43] S. Itoh, H. Okada, K. Koyano, S. Nakamura, Y. Konishi, T. Iwase, and T. Kusaka, "Fetal and neonatal bilirubin metabolism," Frontiers in Pediatrics, vol. 10, 2023.

[44] H. LP, "Neonatal jaundice and liver disease," Neonatal-Perinatal Medicine: Disease of the Fetus and Infant, 1997.

[45] C. Hammerman and M. Kaplan, "Hyperbilirubinemia in the term infant: re-evaluating what we think we know," Clinics in Perinatology, vol. 48, no. 3, pp. 533–554, 2021.

[46] S. Onishi, N. Kawade, S. Itoh, K. Isobe, and S. Sugiyama, "Postnatal development of uridine diphosphate glucuronyltransferase activity towards bilirubin and 2-aminophenol in human liver," Biochemical Journal, vol. 184, no. 3, pp. 705–707, 1979.

[47] M. Kaplan and M. J. Maisels, "Natural history of early neonatal bilirubinemia: a global perspective," Journal of Perinatology, vol. 41, no. 4, pp. 873–878, 2021. [Online]. Available: http://www.nature.com/articles/s41372-020-00901-x

[48] M. J. Maisels, "What's in a name? physiologic and pathologic jaundice: the conundrum of defining normal bilirubin levels in the newborn," Pediatrics, vol. 118, no. 2, pp. 805–807, 2006.

[49] L. Johnson and V. K. Bhutani, "The clinical syndrome of bilirubin-induced neurologic dysfunction," in Seminars in perinatology, vol. 35, no. 3. Elsevier, 2011, pp. 101–113.

[50] S. M. Shapiro, "Definition of the clinical spectrum of kernicterus and bilirubin-induced neurologic dysfunction (BIND)," Journal of perinatology, vol. 25, no. 1, pp. 54–59, 2005.

[51] C. J. Wusthoff and I. M. Loe, "Impact of bilirubin-induced neurologic dysfunction on neurodevelopmental outcomes," Seminars in Fetal and Neonatal Medicine, vol. 20, no. 1, pp. 52–57, 2015.

[52] J. F. Watchko, "Bilirubin-induced neurotoxicity in the preterm neonate," Clinics in perinatology, vol. 43, no. 2, pp. 297–311, 2016.

[53] J. A. Taylor, A. E. Burgos, V. Flaherman, E. K. Chung, E. A. Simpson, N. K. Goyal, I. Von Kohorn, N. Dhepyasuwan, and B. O. through Research for Newborns Network, "Discrepancies between transcutaneous and serum bilirubin measurements," Pediatrics, vol. 135, no. 2, pp. 224–231, 2015.

[54] M. J. Maisels and E. Kring, "Transcutaneous bilirubinometry decreases the need for serum bilirubin measurements and saves money." Pediatrics, vol. 99, no. 4, pp. 599–601, 1997.

[55] V. K. Bhutani, G. R. Gourley, S. Adler, B. Kreamer, C. Dalin, and L. H. Johnson, "Noninvasive measurement of total serum bilirubin in a multiracial predischarge newborn population to assess the risk of severe hyperbilirubinemia," Pediatrics, vol. 106, no. 2, p. e17, 2000. [Online]. Available: https://doi.org/10.1542/peds.106.2.e17

[56] R. E. Schumacher, "Transcutaneous bilirubinometry and diagnostic tests: "the right job for the tool"," Pediatrics, vol. 110, no. 2, pp. 407–408, 2002.

[57] M. J. Maisels, E. M. Ostrea Jr, S. Touch, S. E. Clune, E. Cepeda, E. Kring, K. Gracey, C. Jackson, D. Talbot, and R. Huang, "Evaluation of a new transcutaneous bilirubinometer," Pediatrics, vol. 113, no. 6, pp. 1628–1635, 2004.

[58] S. N. El-Beshbishi, K. E. Shattuck, A. A. Mohammad, and J. R. Petersen, "Hyperbilirubinemia and transcutaneous bilirubinometry," Clinical chemistry, vol. 55, no. 7, pp. 1280–1287, 2009.

[59] C. V. Hulzebos, L. Vitek, C. D. Coda Zabetta, A. Dvořák, P. Schenk, E. A. van der Hagen, C. Cobbaert, and C. Tiribelli, "Screening methods for neonatal hyperbilirubinemia: benefits, limitations, requirements, and novel developments," Pediatric research, vol. 90, no. 2, pp. 272–276, 2021.

[60] M. Mohamed, N. R. Ibrahim, N. Ramli, N. A. Majid, N. M. Yacob, and A. Nasir, "Comparison between the transcutaneous and total serum bilirubin measurement in Malay neonates with neonatal jaundice," The Malaysian Journal of Medical Sciences: MJMS, vol. 29, no. 1, p. 43, 2022.

[61] V. K. Bhutani, G. R. Gourley, S. Adler, B. Kreamer, C. Dalin, and L. H. Johnson, "Noninvasive measurement of total serum bilirubin in a multiracial predischarge newborn population to assess the risk of severe hyperbilirubinemia," Pediatrics, vol. 106, no. 2, pp. e17–e17, 2000.

[62] F. F. Rubaltelli, G. R. Gourley, N. Loskamp, N. Modi, M. Roth-Kleiner, A. Sender, and P. Vert, "Transcutaneous bilirubin measurement: a multicenter evaluation of a new device," Pediatrics, vol. 107, no. 6, pp. 1264–1271, 2001.

[63] G. Nagar, B. Vandermeer, S. Campbell, and M. Kumar, "Reliability of transcutaneous bilirubin devices in preterm infants: a systematic review," Pediatrics, vol. 132, no. 5, pp. 871–881, 2013.

[64] J. Grabenhenrich, L. Grabenhenrich, C. Bührer, and M. Berns, "Transcutaneous bilirubin after phototherapy in term and preterm infants," Pediatrics, vol. 134, no. 5, pp. e1324–e1329, 2014. [Online]. Available: https://doi.org/10.1542/peds.2014-1677

[65] T. Jegathesan, D. M. Campbell, J. G. Ray, V. Shah, H. Berger, R. Z. Hayeems, M. Sgro, and for the NeoHBC, "Transcutaneous versus total serum bilirubin measurements in preterm infants," Neonatology, vol. 118, no. 4, pp. 443–453, 2021. [Online]. Available: https://www.karger.com/Article/FullText/516648

[66] L. Ten Kate, T. van Oorschot, J. Woolderink, S. Teklenburg-Roord, and J. Bekhof, "Transcutaneous bilirubin accuracy before, during, and after phototherapy: A meta-analysis," Pediatrics, vol. 152, no. 6, p. e2023062335, 2023.

[67] V. K. Bhutani, A. R. Stark, L. C. Lazzeroni, R. Poland, G. R. Gourley, S. Kazmierczak, L. Meloy, A. E. Burgos, J. Y. Hall, and D. K. Stevenson, "Predischarge screening for severe neonatal hyperbilirubinemia identifies infants who need phototherapy," The Journal of Pediatrics, vol. 162, no. 3, pp. 477–482.e1, 2013.

[68] R. Cremer, P. Perryman, and D. Richards, "Influence of light on the hyperbilirubinæmia of infants," The Lancet, vol. 271, no. 7030, pp. 1094–1097, 1958.

[69] L. I. Grossweiner, J. B. Grossweiner, and B. Gerald Rogers, "Phototherapy of neonatal jaundice," in The Science of Phototherapy: An Introduction, L. R. Jones, Ed. Berlin/Heidelberg: Springer-Verlag, 2005, pp. 329–335.

[70] M. J. Maisels, "Phototherapy—traditional and nontraditional," Journal of perinatology, vol. 21, no. 1, pp. S93–S97, 2001.

[71] M. J. Maisels and A. F. McDonagh, "Phototherapy for neonatal jaundice," New England Journal of Medicine, vol. 358, no. 9, pp. 920–928, 2008.

[72] V. K. Bhutani, L. H. Johnson, and S. M. Shapiro, "Kernicterus in sick and preterm infants (1999—2002): A need for an effective preventive approach," Seminars in Perinatology, vol. 28, no. 5, pp. 319–325, 2004.

[73] L. M. Gartner, R. N. Snyder, R. S. Chabon, and J. Bernstein, "Kernicterus: high incidence in premature infants with low serum bilirubin concentrations," Pediatrics, vol. 45, no. 6, pp. 906–917, 1970.

[74] M. van de Bor, T. M. van Zeben-van der Aa, S. P. Verloove-Vanhorick, R. Brand, and J. H. Ruys, "Hyperbilirubinemia in preterm infants and neurodevelopmental outcome at 2 years of age: results of a national collaborative survey," Pediatrics, vol. 83, no. 6, pp. 915–920, 1989.

[75] W. Oh, J. E. Tyson, A. A. Fanaroff, B. R. Vohr, R. Perritt, B. J. Stoll, R. A. Ehrenkranz, W. A. Carlo, S. Shankaran, and K. Poole, "Association between peak serum bilirubin and neurodevelopmental outcomes in extremely low birth weight infants," Pediatrics, vol. 112, no. 4, pp. 773–779, 2003.

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

[77] M. Maisels, J. Watchko, V. Bhutani, and D. Stevenson, "An approach to the management of hyperbilirubinemia in the preterm infant less than 35 weeks of gestation," Journal of perinatology, vol. 32, no. 9, pp. 660–664, 2012.

[78] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith et al., "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," Jama, vol. 315, no. 8, pp. 801–810, 2016.

[79] S. A. Coggins and K. Glaser, "Updates in late-onset sepsis: risk assessment, therapy, and outcomes," Neoreviews, vol. 23, no. 11, pp. 738–755, 2022.

[80] R. Hayes, J. Hartnett, G. Semova, C. Murray, K. Murphy, L. Carroll, H. Plapp, L. Hession, J. O'Toole, D. McCollum et al., "Neonatal sepsis definitions from randomised clinical trials," Pediatric research, vol. 93, no. 5, pp. 1141–1148, 2023.

[81] A. C. Seale, H. Blencowe, A. A. Manu, H. Nair, R. Bahl, S. A. Qazi, A. K. Zaidi, J. A. Berkley, S. N. Cousens, and J. E. Lawn, "Estimates of possible severe bacterial infection in neonates in sub-Saharan Africa, south Asia, and Latin America for 2012: a systematic review and meta-analysis," The Lancet infectious diseases, vol. 14, no. 8, pp. 731–741, 2014.

[82] A. L. Shane, P. J. Sánchez, and B. J. Stoll, "Neonatal sepsis," The lancet, vol. 390, no. 10104, pp. 1770–1780, 2017.

[83] I. H. Celik, M. Hanna, F. E. Canpolat, and M. Pammi, "Diagnosis of neonatal sepsis: the past, present and future," Pediatric research, vol. 91, no. 2, pp. 337–350, 2022.

[84] B. J. Stoll, N. I. Hansen, E. F. Bell, S. Shankaran, A. R. Laptook, M. C. Walsh, E. C. Hale, N. S. Newman, K. Schibler, W. A. Carlo et al., "Neonatal outcomes of extremely preterm infants from the NICHD neonatal research network," Pediatrics, vol. 126, no. 3, pp. 443–456, 2010.

[85] J. L. Wynn, "Defining neonatal sepsis," Current opinion in pediatrics, vol. 28, no. 2, pp. 135–140, 2016.

[86] M. Dahou, M. Lehlimi, Z. Korchi, R. Chaini, A. Badre, M. Chemsi, and A. Habzi, "Early-onset neonatal sepsis: The challenges of management," American Journal of Pediatrics, vol. 10, no. 3, pp. 34–40, 2024.

[87] B. J. Stoll, N. Hansen, A. A. Fanaroff, L. L. Wright, W. A. Carlo, R. A. Ehrenkranz, J. A. Lemons, E. F. Donovan, A. R. Stark, J. E. Tyson et al., "Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network," Pediatrics, vol. 110, no. 2, pp. 285–291, 2002.

[88] R. G. Greenberg, S. Kandefer, B. T. Do, P. B. Smith, B. J. Stoll, E. F. Bell, W. A. Carlo, A. R. Laptook, P. J. Sánchez, S. Shankaran et al., "Late-onset sepsis in extremely premature infants: 2000–2011," The Pediatric infectious disease journal, vol. 36, no. 8, pp. 774–779, 2017.

[89] M. A. Verboon-Maciolek, S. F. Thijsen, M. A. Hemels, M. Menses, A. M. van Loon, T. G. Krediet, L. J. Gerards, A. Fleer, H. A. Voorbij, and G. T. Rijkers, "Inflammatory mediators for the diagnosis and treatment of sepsis in early infancy," Pediatric research, vol. 59, no. 3, pp. 457–461, 2006.

[90] P.-Y. Iroh Tam and C. M. Bendel, "Diagnostics for neonatal sepsis: current approaches and future directions," Pediatric research, vol. 82, no. 4, pp. 574–583, 2017.

[91] B. Y. H. Wee, J. H. Lee, Y. H. Mok, and S.-L. Chong, "A narrative review of heart rate and variability in sepsis," Annals of translational medicine, vol. 8, no. 12, 2020.

[92] K. Guerti, H. Devos, M. M. Ieven, and L. M. Mahieu, "Time to positivity of neonatal blood cultures: fast and furious?" Journal of medical microbiology, vol. 60, no. 4, pp. 446–453, 2011.

[93] F. Brunkhorst, U. Heinz, and Z. Forycki, "Kinetics of procalcitonin in iatrogenic sepsis," Intensive care medicine, vol. 24, no. 8, pp. 888–889, 1998.

[94] M. Gilfillan and V. Bhandari, "Biomarkers for the diagnosis of neonatal sepsis and necrotizing enterocolitis: Clinical practice guidelines," Early human development, vol. 105, pp. 25–33, 2017.

[95] J. Eichberger, E. Resch, and B. Resch, "Diagnosis of neonatal sepsis: the role of inflammatory markers," Frontiers in Pediatrics, vol. 10, p. 840288, 2022.

[96] E. Kocabaş, A. Sarikçioğlu, N. Aksaray, G. Seydaoğlu, Y. Seyhun, and A. Yaman, "Role of procalcitonin, C-reactive protein, interleukin-6, interleukin-8 and tumor necrosis factor-alpha in the diagnosis of neonatal sepsis," The Turkish journal of pediatrics, vol. 49, no. 1, pp. 7–20, 2007.

[97] H. Küster, M. Weiss, A. E. Willeitner, S. Detlefsen, I. Jeremias, J. Zbojan, R. Geiger, G. Lipowsky, and G. Simbruner, "Interleukin-1 receptor antagonist and interleukin-6 for early diagnosis of neonatal sepsis 2 days before clinical manifestation," The Lancet, vol. 352, no. 9136, pp. 1271–1277, 1998.

[98] L. Dillenseger, C. Langlet, S. Iacobelli, T. Lavaux, C. Ratomponirina, M. Labenne, D. Astruc, F. Severac, J. B. Gouyon, and P. Kuhn, "Early inflammatory markers for the diagnosis of late-onset sepsis in neonates: the nosodiag study," Frontiers in pediatrics, vol. 6, p. 346, 2018.

[99] P. Ng, S. Cheng, K. Chui, T. Fok, M. Wong, W. Wong, R. Wong, and K. Cheung, "Diagnosis of late onset neonatal sepsis with cytokines, adhesion molecule, and C-reactive protein in preterm very low birthweight infants," Archives of Disease in Childhood-Fetal and Neonatal Edition, vol. 77, no. 3, pp. F221–F227, 1997.

[100] K. D. Fairchild and T. M. O'Shea, "Heart rate characteristics: physiomarkers for detection of late-onset neonatal sepsis," Clinics in perinatology, vol. 37, no. 3, pp. 581–598, 2010.

[101] R. P. Dellinger, M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke et al., "Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012," Critical care medicine, vol. 41, no. 2, pp. 580–637, 2013.

[102] J.-L. Vincent, "The clinical challenge of sepsis identification and monitoring," PLoS medicine, vol. 13, no. 5, p. e1002022, 2016.

[103] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg et al., "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," Critical care medicine, vol. 34, no. 6, pp. 1589–1596, 2006.

[104] R. A. Polin, Committee on Fetus and Newborn, L.-A. Papile, J. E. Baley, V. K. Bhutani, W. A. Carlo, J. Cummings, P. Kumar, R. C. Tan, K. S. Wang et al., "Management of neonates with suspected or proven early-onset bacterial sepsis," Pediatrics, vol. 129, no. 5, pp. 1006–1015, 2012.

[105] K. M. Puopolo, W. E. Benitz, T. E. Zaoutis, J. Cummings, S. Juul, I. Hand, E. Eichenwald, B. Poindexter, D. L. Stewart, S. W. Aucott et al., "Management of neonates born at $\leq$34 6/7 weeks' gestation with suspected or proven early-onset bacterial sepsis," Pediatrics, vol. 142, no. 6, 2018.

[106] ——, "Management of neonates born at $\geq$35 0/7 weeks' gestation with suspected or proven early-onset bacterial sepsis," Pediatrics, vol. 142, no. 6, 2018.

[107] The Royal Children's Hospital Melbourne, "Sepsis —assessment and management," https://www.rch.org.au/clinicalguide/guideline_index/SEPSIS_assessment_and_management/, 2020, Accessed: 2024-09-18.

[108] NICE, "Neonatal infection: antibiotics for prevention and treatment," https://www.nice.org.uk/guidance/ng195, 2021, Accessed: 2024-09-18.

[109] P. Ng, "Diagnostic markers of infection in neonates," Archives of Disease in Childhood-Fetal and Neonatal Edition, vol. 89, no. 3, pp. F229–F235, 2004.

[110] V. S. Kuppala, J. Meinzen-Derr, A. L. Morrow, and K. R. Schibler, "Prolonged initial empirical antibiotic treatment is associated with adverse outcomes in premature infants," The Journal of pediatrics, vol. 159, no. 5, pp. 720–725, 2011.

[111] R. Singh, L. Sripada, and R. Singh, "Side effects of antibiotics during bacterial infection: mitochondria, the main target in host cell," Mitochondrion, vol. 16, pp. 50–54, 2014.

[112] A. Zea-Vera and T. J. Ochoa, "Challenges in the diagnosis and management of neonatal sepsis," Journal of tropical pediatrics, vol. 61, no. 1, pp. 1–13, 2015.

[113] World Health Organization, "Antimicrobial resistance," https://www.who.int/en/news-room/fact-sheets/detail/antimicrobial-resistance, November 2023, Accessed: 2024-09-30.

[114] I. Sim, P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann, and P. C. Tang, "Clinical decision support systems for the practice of evidence-based medicine," Journal of the American Medical Informatics Association, vol. 8, no. 6, pp. 527–534, 2001.

[115] J. A. Osheroff, J. Teich, D. Levick, L. Saldana, F. Velasco, D. Sittig, K. Rogers, and R. Jenders, Improving outcomes with clinical decision support: an implementer's guide. Himss Publishing, 2012.

[116] K. Cresswell, A. Majeed, D. Bates, and A. Sheikh, "Computerised decision support systems for healthcare professionals: an interpretative review," Informatics in primary care, vol. 20, no. 2, pp. 115–128, 2012.

[117] A. M. Pereira, C. Jácome, R. Amaral, T. Jacinto, and J. A. Fonseca, "Real-time clinical decision support at the point of care," in Implementing Precision Medicine in Best Practices of Chronic Airway Diseases. Elsevier, 2019, pp. 125–133.

[118] A. Rao and J. Palma, "Clinical decision support in the neonatal ICU," Seminars in Fetal and Neonatal Medicine, vol. 27, no. 5, p. 101332, 2022.

[119] O. Bodenreider, R. Cornet, and D. J. Vreeman, "Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm," Yearbook of medical informatics, vol. 27, no. 01, pp. 129–139, 2018.

[120] J. Kabachinski, "What is health level 7?" Biomedical Instrumentation & Technology, vol. 40, no. 5, pp. 375–379, 2006.

[121] K. Kawamoto, G. Del Fiol, D. F. Lobach, and R. A. Jenders, "Standards for scalable clinical decision support: need, current and emerging standards, gaps, and proposal for progress," The open medical informatics journal, vol. 4, p. 235, 2010.

[122] J. S. Malak, H. Zeraati, F. S. Nayeri, R. Safdari, and A. D. Shahraki, "Neonatal intensive care decision support systems using artificial intelligence techniques: a systematic review," Artificial Intelligence Review, vol. 52, pp. 2685–2704, 2019.

[123] M. A. Beyer, "The importance of big data: A definition," 2012.

[124] S. Ramgopal, L. N. Sanchez-Pinto, C. M. Horvat, M. S. Carroll, Y. Luo, and T. A. Florin, "Artificial intelligence-based clinical decision support in pediatrics," Pediatric research, vol. 93, no. 2, pp. 334–341, 2023.

[125] K. Beam, P. Sharma, P. Levy, and A. L. Beam, "Artificial intelligence in the neonatal intensive care unit: the time is now," Journal of Perinatology, vol. 44, no. 1, pp. 131–135, 2024.

[126] A. A. Vinks, N. C. Punt, F. Menke, E. Kirkendall, D. Butler, T. J. Duggan, D. E. Cortezzo, S. Kiger, T. Dietrich, P. Spencer et al., "Electronic health record–embedded decision support platform for morphine precision dosing in neonates," Clinical Pharmacology & Therapeutics, vol. 107, no. 1, pp. 186–194, 2020.

[127] A. Frymoyer, C. Stockmann, A. L. Hersh, S. Goswami, and R. J. Keizer, "Individualized empiric vancomycin dosing in neonates using a model-based approach," Journal of the Pediatric Infectious Diseases Society, vol. 8, no. 2, pp. 97–104, 2019.

[128] B. Rittmann and M. P. Stevens, "Clinical decision support systems and their role in antibiotic stewardship: a systematic review," Current infectious disease reports, vol. 21, pp. 1–12, 2019.

[129] M. Beaudoin, F. Kabanza, V. Nault, and L. Valiquette, "An antimicrobial prescription surveillance system that learns from experience," AI Magazine, vol. 35, no. 1, pp. 15–15, 2014.

[130] ——, "Evaluation of a machine learning capability for a clinical decision support system to enhance antimicrobial stewardship programs," Artificial intelligence in medicine, vol. 68, pp. 29–36, 2016.

[131] B.-H. Tang, B.-F. Yao, W. Zhang, X.-F. Zhang, S.-M. Fu, G.-X. Hao, Y. Zhou, D.-Q. Sun, G. Liu, J. van den Anker et al., "Optimal use of β-lactams in neonates: machine learning-based clinical decision support system," EBioMedicine, vol. 105, 2024.

[132] Stuart Turner, BiliTool Inc., "BiliTool^{TM}," https://bilitool.org/index.php, 2024, Accessed: 2024-08-22.

[133] J. D. Petersen, M. Lozovatsky, D. Markovic, R. Duncan, S. Zheng, A. Shamsian, S. Kagele, and M. K. Ross, "Clinical decision support for hyperbilirubinemia risk assessment in the electronic health record," Academic pediatrics, vol. 20, no. 6, pp. 857–862, 2020.

[134] J. P. Palma and Y. H. Arain, "Development of a web-based decision support tool to operationalize and optimize management of hyperbilirubinemia in preterm infants." Clinics in Perinatology, vol. 43, no. 2, pp. 375–383, 2016.

[135] Y. Arain, J. M. Banda, J. Faulkenberry, V. K. Bhutani, and J. P. Palma, "Clinical decision support tool for phototherapy initiation in preterm infants," Journal of Perinatology, vol. 40, no. 10, pp. 1518–1523, 2020.

[136] M. Gomez, C. Bielza, J. A. Fernández del Pozo, and S. Rios-Insua, "A graphical decision-theoretic model for neonatal jaundice," Medical Decision Making, vol. 27, no. 3, pp. 250–265, 2007.

[137] G. J. Escobar, K. M. Puopolo, S. Wi, B. J. Turk, M. W. Kuzniewicz, E. M. Walsh, T. B. Newman, J. Zupancic, E. Lieberman, and D. Draper, "Stratification of risk of early-onset sepsis in newborns $\geq$34 weeks' gestation," Pediatrics, vol. 133, no. 1, pp. 30–36, 2014.

[138] K. J. Pettinger, K. Mayers, L. McKechnie, and B. Phillips, "Sensitivity of the kaiser permanente early-onset sepsis calculator: a systematic review and meta-analysis," EClinicalMedicine, vol. 19, 2020.

[139] S. Mani, A. Ozdas, C. Aliferis, H. A. Varol, Q. Chen, R. Carnevale, Y. Chen, J. Romano-Keeler, H. Nian, and J.-H. Weitkamp, "Medical decision support using machine learning for early detection of late-onset neonatal sepsis," Journal of the American Medical Informatics Association, vol. 21, no. 2, pp. 326–336, 2014.

[140] A. J. Masino, M. C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C. P. Bonafide, F. Balamuth, M. Schmatz, and R. W. Grundmeier, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," PloS one, vol. 14, no. 2, p. e0212665, 2019.

[141] E. Persad, K. Jost, A. Honoré, D. Forsberg, K. Coste, H. Olsson, S. Rautiainen, and E. Herlenius, "Neonatal sepsis prediction through clinical decision support algorithms: a systematic review," Acta Paediatrica, vol. 110, no. 12, pp. 3201–3226, 2021.

[142] K. D. Fairchild and J. L. Aschner, "HeRO monitoring to reduce mortality in NICU patients," Research and Reports in Neonatology, pp. 65–76, 2012.

[143] M. P. Griffin, T. M. O'Shea, E. A. Bissonette, F. E. Harrell, D. E. Lake, and J. R. Moorman, "Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness," Pediatric research, vol. 53, no. 6, pp. 920–926, 2003.

[144] J. R. Moorman, W. A. Carlo, J. Kattwinkel, R. L. Schelonka, P. J. Porcelli, C. T. Navarrete, E. Bancalari, J. L. Aschner, M. W. Walker, J. A. Perez et al., "Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial," The Journal of pediatrics, vol. 159, no. 6, pp. 900–906, 2011.

[145] S. Gómez-Quintana, C. E. Schwarz, I. Shelevytsky, V. Shelevytska, O. Semenova, A. Factor, E. Popovici, and A. Temko, "A framework for AI-assisted detection of patent ductus arteriosus from neonatal phonocardiogram," in Healthcare, vol. 9, no. 2. MDPI, 2021, p. 169.

[146] A. Mikhno and C. M. Ennett, "Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database," in 2012 Annual international conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2012, pp. 5094–5097.

[147] A. Temko, W. Marnane, G. Boylan, and G. Lightbody, "Clinical implementation of a neonatal seizure detection algorithm," Decision support systems, vol. 70, pp. 86–96, 2015.

[148] E. Low, N. Stevenson, A. Temko, G. Lightbody, W. Marnane, V. Livingstone, S. Mathieson, C. Ryan, J. Rennie, and G. Boylan, "Clinical validation of a neonatal seizure detection algorithm," Pediatric Research, vol. 70, no. 5, pp. 135–135, 2011.

[149] S. A. Raurale, G. B. Boylan, S. R. Mathieson, W. P. Marnane, G. Lightbody, and J. M. O'Toole, "Grading hypoxic-ischemic encephalopathy in neonatal EEG with convolutional neural networks and quadratic time-frequency distributions," Journal of Neural Engineering, vol. 18, no. 4, p. 046007, 2021.

[150] J. S. Gilchrist, "Clinical decision support system using real-time data analysis for a neonatal intensive care unit," Ph.D. dissertation, Carleton University, 2012.

[151] B. Norgeot, G. Quer, B. K. Beaulieu-Jones, A. Torkamani, R. Dias, M. Gianfrancesco, R. Arnaout, I. S. Kohane, S. Saria, E. Topol et al., "Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist," Nature medicine, vol. 26, no. 9, pp. 1320–1324, 2020.

[152] M. Frize, E. Bariciak, and S. Weyand, "Suggested criteria for successful deployment of a clinical decision support system (CDSS)," in 2010 IEEE International Workshop on Medical Measurements and Applications. IEEE, 2010, pp. 69–72.

[153] E. S. Berner and T. J. La Lande, "Overview of clinical decision support systems," Clinical decision support systems: Theory and practice, pp. 1–17, 2016.

[154] P. Van Schaik, D. Flynn, A. Van Wersch, A. Douglass, and P. Cann, "The acceptance of a computerised decision-support system in primary care: A preliminary investigation," Behaviour & information technology, vol. 23, no. 5, pp. 321–326, 2004.

[155] E. H. Shortliffe and M. J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," Jama, vol. 320, no. 21, pp. 2199–2200, 2018.

[156] R. Shibl, M. Lawley, and J. Debuse, "Factors influencing decision support system acceptance," Decision Support Systems, vol. 54, no. 2, pp. 953–961, 2013.

[157] N. Bressan, A. James, and C. McGregor, "Trends and opportunities for integrated real time neonatal clinical decision support," in Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics. IEEE, 2012, pp. 687–690.

[158] A. R. Spitzer, D. Ellsbury, and R. H. Clark, "The Pediatrix BabySteps® Data Warehouse—a unique national resource for improving outcomes for neonates," The Indian Journal of Pediatrics, vol. 82, pp. 71–79, 2015.

[159] E. Meyeroff and P. Tremoulet, "Etiometry's T3 heuristic evaluation," in Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, vol. 10, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2021, pp. 37–41.

[160] A. Asfari, "Artificial intelligence role and clinical decision support system extubation readiness trail and Etiometry scoring system," Biomedical Journal of Scientific & Technical Research, vol. 35, no. 1, pp. 27 291–27 293, 2021.

[161] H. Singh, R. Mallaiah, G. Yadav, N. Verma, A. Sawhney, and S. K. Brahmachari, "iCHRCloud: web & mobile based child health imprints for smart healthcare," Journal of medical systems, vol. 42, no. 1, p. 14, 2018.

[162] H. Singh, R. Kaur, A. Gangadharan, A. K. Pandey, A. Manur, Y. Sun, S. Saluja, S. Gupta, J. P. Palma, and P. Kumar, "Neo-bedside monitoring device for integrated neonatal intensive care unit (iNICU)," IEEE Access, vol. 7, pp. 7803–7813, 2018.

[163] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," IEEE access, vol. 6, pp. 52 138–52 160, 2018.

[164] A. Honoré, D. Forsberg, K. Adolphson, S. Chatterjee, K. Jost, and E. Herlenius, "Vital sign-based detection of sepsis in neonates using machine learning," Acta Paediatrica, vol. 112, no. 4, pp. 686–696, 2023.

[165] X. Li, X. Xu, F. Xie, X. Xu, Y. Sun, X. Liu, X. Jia, Y. Kang, L. Xie, F. Wang et al., "A time-phased machine learning model for real-time prediction of sepsis in critical care," Critical Care Medicine, vol. 48, no. 10, pp. e884–e888, 2020.

[166] H.-F. Deng, M.-W. Sun, Y. Wang, J. Zeng, T. Yuan, T. Li, D.-H. Li, W. Chen, P. Zhou, Q. Wang et al., "Evaluating machine learning models for sepsis prediction: A systematic review of methodologies," Iscience, vol. 25, no. 1, 2022.

CHAPTER

2 Methods and Tools

This dissertation is conducted in the context of a prospective clinical project of CARESS-Premi that aims to develop computer-assisted diagnostic tools to help clinicians improve neonatal care and outcomes in the Neonatal intensive care units (NICU) settings. Based on the CARESS-Premi project, a database was established, which was used for all subsequent studies presented in the dissertation. In this chapter, we begin with an introduction to the CARESS-Premi project, in which the clinical protocols, data acquisition system and general information of the CARESS-Premi database are mentioned. Then we introduce a proposed complete data processing pipeline from raw signal detection to feature extraction, which serves as a fundamental methodological framework that runs across multiple studies in the dissertation. The following sections present the statistical techniques, machine learning algorithms and explainability analyses, which were applied in different studies throughout the dissertation.

2.1 The CARESS-Premi Project

In this section, we introduce the CARESS-Premi clinical protocol and describe a cloud-based system that was employed for data acquisition and transmission in the project. The overview and some illustrations of the CARESS-Premi database are then presented.

2.1.1 The CARESS-Premi study

The CARESS-Premi clinical study, "Contribution of Real Time Analyses of CARdio-RESpiratory Signals to the Diagnosis of Infection in PREterM Infants", is a prospective national multi-center observational cohort study with blinded analysis by healthcare staff. This study was registered on Clinicaltrials.gov (NCT01611740), it concerns approximately 500 preterm infants recruited between October 2012 and November 2018 across the Neonatal intensive care units (NICU) in three French hospitals: the University Hospital Centers of Rennes (*CHU Rennes*), Lille and Angers. The local ethics committee (CPP Ouest V Rennes) approved the study (protocol number: 11/14-803) and the informed parental consent of the associated children was obtained.

The project was proposed with the fundamental perspective of developing new computer-assisted diagnostic tools, i.e., Clinical decision support systems (CDSS), to help caregivers in the early diagnosis and intervention of nosocomial infection, both non-invasively and continuously, at the bedside of preterm infants. Achieving this could lead to a reduction in the need for blood

sampling for the diagnosis of infection, pain for the vulnerable population, duration of hospitalization, risks of related complications and neurodevelopmental damages, and overall healthcare costs. Furthermore, this could open the possibility of creating such CDSS which can be adapted to telemedicine practice.

The project aims to create innovative tools in a combination of available clinical data and multi-signal analysis, including cardiac cycle duration variability, respiratory cycle amplitude and duration variability, Photoplethysmography signals and their relationships. These advanced tools would be integrated into a continuous monitoring system, and preliminarily evaluated in a real-time clinical context. Furthermore, by obtaining longitudinal physiological data correlated to clinical status and multiple underlying physiological processes in an unselected cohort, this system seeks to enhance our understanding of perinatal health status and development, ultimately improving the preterm infants' outcomes and the quality of neonatal care.

The inclusion in the study was prospective with all preterm newborns arriving in the departments potentially included in the protocol before 4 days of postnatal age.

The inclusion criteria for the study subjects were:

- ✓ a preterm birth between 24 and 32 weeks of gestation;
- ✓ a birth weight >500 grams;
- ✓ a postnatal age >3 days;
- ✓ a postmenstrual age <34 weeks;
- ✓ parental information and collection of non-opposition.

The exclusion criteria included:

- x a malformation syndrome;
- x a severe neurological lesion (grade 4 Intraventricular hemorrhage (IVH), cavitary periventricular leukomalacia, post-perinatal ischemia).

During the follow-ups, the data collection involves three parts: *i)* Clinical data (not specific to the study) were collected from patient records every 6 hours in a relational event database, such as parameters and configurations of ventilatory supports, temperature, blood pressure, pain-comfort scores, medical treatments and so forth. The quality of the data collection was ensured by a dedicated clinical research nurse at each center. *ii)* Laboratory data collection was performed in the hospital's electronic database. *iii)* The collection of cardio-respiratory signals was prospective. All included infants were continuously monitored for their cardio-respiratory status from the beginning to the end of the inclusion periods (IntelliVue MP40 Philips Medical System, Eindhoven, Netherlands). Signals, including Electrocardiogram (ECG), transthoracic impedance ventilation and plethysmography, were acquired from monitoring systems, de-identified, transferred and stored progressively using a prototype cloud-based system named ASCENT developed by the team (*SEPIA* of *LTSI - INSERM U1099*), which is introduced as follows.

2.1.2 Cloud-based anonymized system for clinical experimentation (ASCENT)

In Neonatal intensive care units (NICU), preterm infants are connected to monitoring devices that continuously monitor and acquire a comprehensive set of clinical data. These data include not only cardio-respiratory signals such as ECG and transthoracic impedance ventilation but also various types of alarms and textual data.

Cardio-respiratory monitors are typically linked to a central station that can store data segments. Most of these systems generate low-resolution, partially-sampled signal segments which are stored on closed, vendor-specific formats and databases, making their exploitation difficult for massive signal processing and analysis. Our team (*SEPIA* in *LTSI - INSERM U1099*) has developed a stand-alone application called SYNAPSE, capable of connecting to these monitors and implementing vendor-specific communication protocols to acquire the sensed main clinical data with the highest possible resolution. The SYNAPSE application can execute on a dedicated server station, enabling real-time data acquisition from up to eight monitors simultaneously. Another version of SYNAPSE has been developed for embedded systems and has been deployed using Raspberry Pi modules connected to each monitor. In the context of clinical research protocols, SYNAPSE applications run continuously, acquiring all the available data from preterm infants, from their arrival at the NICU until their discharge. All the data acquired by the SYNAPSE applications are stored locally as a set of files containing 30-minute segments of all signals and associated clinical data.

Aiming to facilitate the sensing, storage, transfer and centralized data analytics of longitudinal clinical data, acquired from heterogeneous medical devices from multiple hospitals, a system named ASCENT (Anonymised System for Clinical experimENTation) was thus proposed in the context of clinical research protocols, in our case, the CARESS-Premi protocol. As demonstrated in Figure 2.1, the system relies on four main components: *1)* the AscentLive component, *2)* the web server component, *3)* the data server component and *4)* the data analytics component. The "AscentLive" is a light client application for secured, de-identified and asynchronous data transfer, also developed by *LTSI - INSERM U1099*. The web server component offers user authentication and is the main interface of the ASCENT system. The data server component handles data storage and file management issues. The data analytics component is able to call protocol-specific workflows, defined by several scripts, in order to process the raw data, perform feature extraction and apply machine-learning methods, contributing to the objectives of each clinical protocol handled by the ASCENT system.

Figure 2.1 illustrates the general architecture of ASCENT system. The left panel of the architecture presents the infrastructure in the clinical space. One or more clinical centers can be involved in patient inclusion. Data acquisition from the monitors is performed for each patient using the SYNAPSE applications that store the data locally. Then local anonymization and encryption of the raw data are performed by the "AscentLive" applications. Subsequently, a clinical investigator uploads the acquired data to the web server in the private cloud space, after secure authentication on

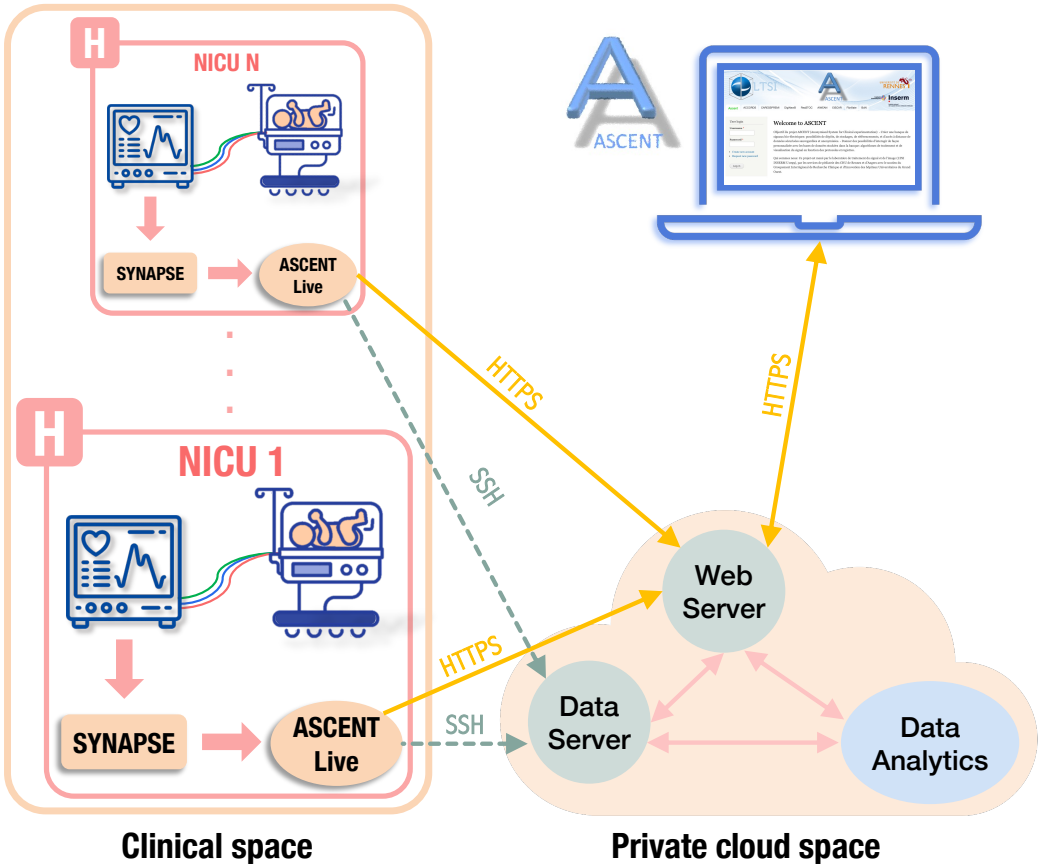

Figure 2.1: General architecture of could-based ASCENT system.

the ASCENT server. The investigator can also initiate a data transfer procedure to the private cloud by connecting to the ASCENT web interface using a standard web browser via HTTPS protocol. Data processing and analysis are performed on the servers in a distributed fashion using the cluster facilities in the private cloud space (right panel). The analysis results are available to specific authenticated users, depending on their specific rights, via a standard web HTTPS connection.

2.1.3 The CARESS-Premi database

Thanks to the ASCENT system (Section 2.1.2), the data acquisition and retrieval for the CARESS-Premi study was made possible. Through the collaborative efforts of clinical and research teams, the database was finalized by July 2021. We introduce the CARESS-Premi database from three aspects: data composition and volume, data compilation process, and presenting some snippets of the monitoring signals.

Data composition and volume

Within the framework of the CARESS-Premi clinical project, as of the end of data collection (February 2021), a database containing 519 patients was established, including 414 patients hospitalized at the Rennes clinical center (dominant center), 42 patients and 63 patients at the Lille and Angers centers, respectively.

In the entire database, 9 (or 11) major deviations from the patient selection criteria were reviewed and verified, resulting in a total of 510 evaluable subjects in the database. These deviations are as follows:

- x 3 with malformation syndromes: right ventricular hypoplasia, diaphragmatic hernia and Prader-Willi Syndrome;
- x 5 with severe neurological lesions: grade 4 IVH, hypoxic-ischemic encephalopathy and periventricular leukomalacia;
- x 1 with mixed conditions: septal agenesis;
- \- 2 late neurological lesions (could be excluded from specific studies as appropriate).

Regarding the data volume, the CARESS-Premi database consists of two main partitions: tabular data for clinical information and time series data for physiological signals.

For tabular data, an Excel file with 10 sheets was created, covering basic demographic information of 519 patients, 4,760 observations on respiratory data, 560 observations about transfusion information, 2,209 observations about jaundice, 1,739 observations about pain management, 7,090 observations about skin-to-skin information, 363 observations about CRP >5 mg/L, 454 observations about clinical events, 233 observations about antibiotics and a supplementary table for other information, totaling more than 200,000 rows and 1.8 million entries.

For the time series data, although some of the initially enrolled patients had missing monitoring signals or data in bad formats before the acquisition and processing system (ASCENT) was stabilized, a massive amount of data was obtained. Numerically, during the CARESS-Premi project, there were over 2,600 times of data submissions uploaded from the ASCENT by the involved clinical investigator, totaling more than 2.6 million compressed binary files (each containing 30 minutes of electrophysiological signals) and occupying approximately 1 TB of storage space. If we add this up, it is equivalent to about 30 years of monitoring signals, which averages out to about 20 days of signal acquisition per patient.

Data compilation

In the CARESS-Premi project, the cardio-respiratory data segments were continuously collected from multiple clinical sites (details refer to Section 2.1.2), and these 30-minute data segments were processed, compressed and saved as zipped binary files (with the file extension of *bin.gz*). In order to make use of the raw data for further analysis, a compilation procedure was designed to

extract the raw electrophysiological signal from binary files and compile it into a Hierarchical data format (HDF5) file (with the file extension of *.h5*).

Each of the 30-minute binary files includes five channels of recordings: ECG signals from leads I, II and III with a sampling frequency of 500 Hz, a plethysmographic recording with a sampling frequency of 125 Hz, and a transthoracic impedance estimation of respiratory activity with a sampling frequency of 62.5 Hz.

For extraction of certain data segments, a study-specific metadata file in *.csv* format is prepared. The metadata is usually extracted from the study's electronic health records, i.e., the tabular data described in Section 2.1.3. Each line of the metadata file should correspond to a targeted event observation, such as bilirubin measurement and sepsis annotation, as well as relevant clinical information such as PNA and birth weight required for the study.

The compilation process is demonstrated in Figure 2.2. For each interested clinical event (*TargetedEvent*), one *.h5* file is created. The patient ID (*PatientID*) and timestamp (*EventDateTime*) of this event are used as references to query, sort and select all raw signal files across all monitoring channels in the database. Then, using the event instant as an anchor, the associated signal with a predefined duration before and after (customized lengths) the event is thus selected and saved as a dataset in this *.h5* file under the *Signal* group. The *Metadata* group is also added to the *.h5* file to annotate this extraction and document the information from the metadata file. The structure of the final *.h5* file is shown in the middle left panel of Figure 2.2. In addition, a complied metadata *.h5* file summarizing all targeted events available in the database for each patient is generated, with the structure shown in the middle right panel of Figure 2.2.

Monitoring signals

Shown in Figure 2.3 are signal snippets from two random patients in the database, giving examples of how the monitored waveforms in five channels look like, displayed in two resolutions: 24-second and 6-second windows. The first three columns present the 3-lead ECG signals in microvolts, and it can be noticed that the monitoring waveform was recorded in only one channel (one ECG lead) at a time. During the example intervals, the valid channel for the first patient (patient 01017) is ECG-III, while it is channel ECG-II for the second patient (patient 01413). Moreover, effective leads providing reliable data or channels actively connected to the babies by the caregivers can alternate between the leads during such long-term and continuous monitoring. This variability can be influenced by unpredictable, sporadic technical or human interference such as interventions by medical personnel to reposition or adjust electrodes, parental interactions (e.g., skin-to-skin contact and feeding), or infant movement, etc. The other two columns display plethysmographic recordings and respiratory activity estimations in Ohms via transthoracic impedance estimation of respiratory activity.

In the dissertation, we particularly focus on the ECG signals of very preterm infants (GA be-

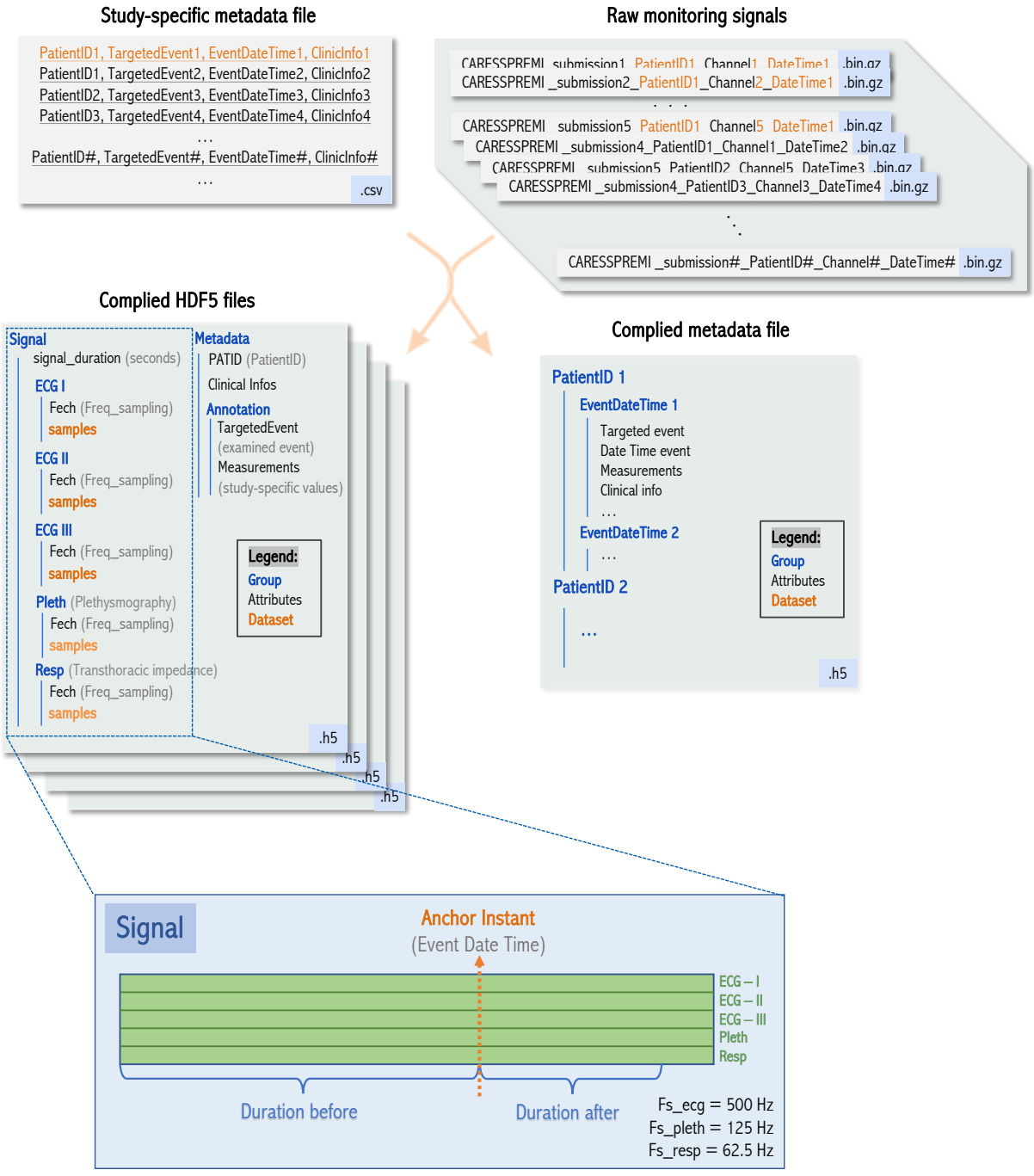

Figure 2.2: Data compilation process of CARESS-Premi database.

tween 24 and 32 weeks), aiming to analyze the variations in the length and amplitude of the cardiac cycle and their potential to improve the diagnostic performance in different neonatal conditions in daily clinical practice in the NICU scenarios.

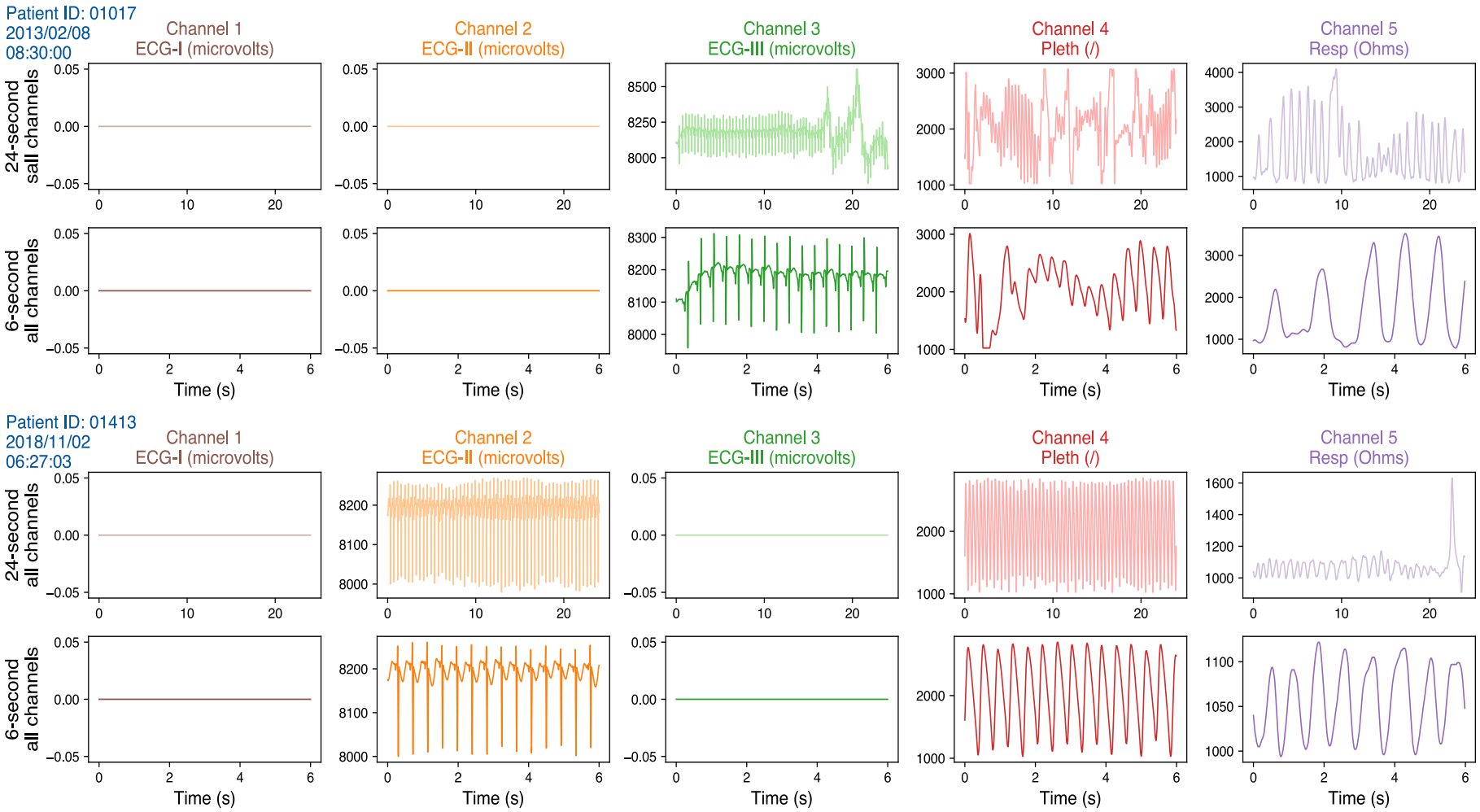

Figure 2.3: Examples for monitoring signals in five channels of CARESS-Premi database.

2.2 Cardiac Signal Processing Pipeline

This section is dedicated to proposing an automatic processing chain integrating raw ECG signal processing, QRS detection, time-series denoising, stationary analysis and HRV analysis. This pipeline accepts cardiac monitoring signals from real clinical practice and produces optimal stationary segments and utile parameters for analysis while greatly reducing the labor and time involved. An overall workflow of the proposed processing chain is described in Figure 2.4. Apart from the first (details refer to Section 2.1.3) and last stages (depends on the studies, see Section 2.3 and Section 2.4), the main body of cardiac signal processing consists of five major steps:

1. Evaluation of ECG signal quality and selection of the good channel;
2. Detection of QRS complex and extraction of RR series;
3. Correction of artifacts in obtained RR sequences;
4. Analysis of time series stationarity;
5. Analysis of HRV characteristics.

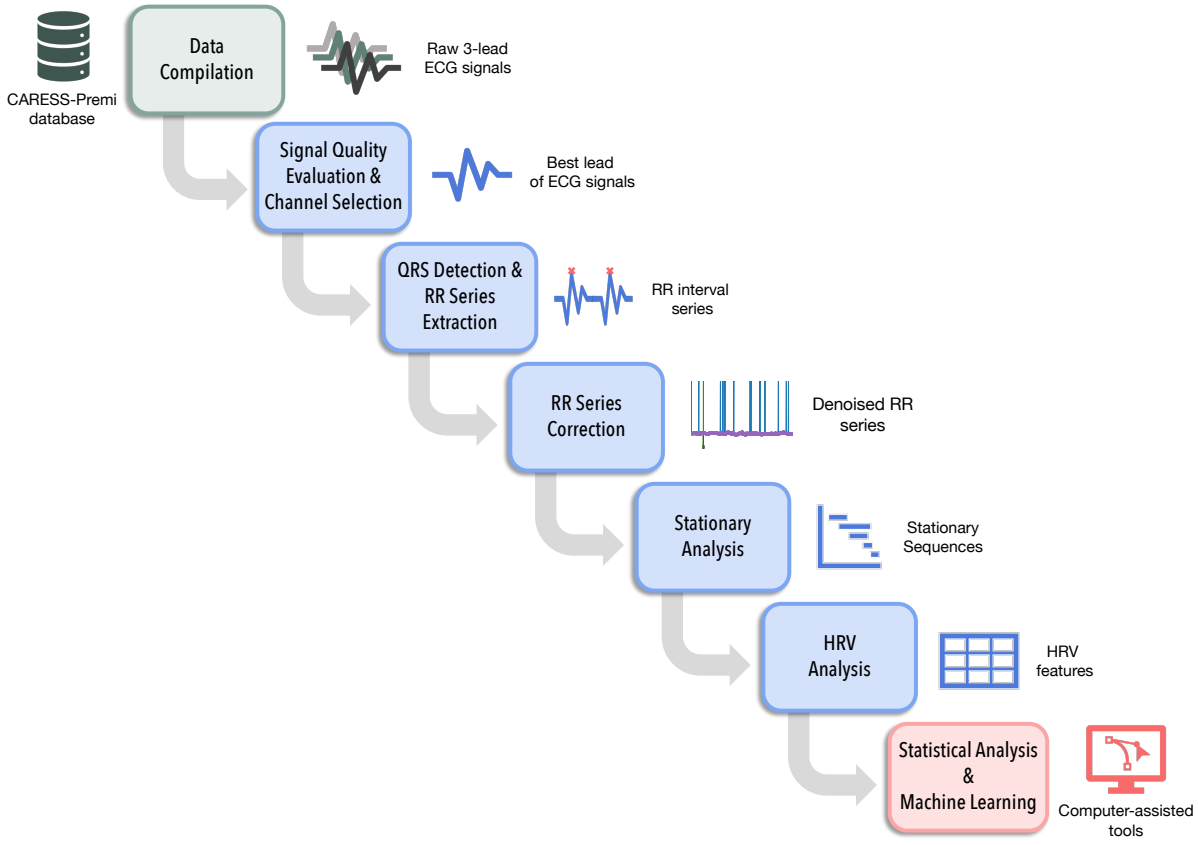

Figure 2.4: Illustration of different stages of the proposed cardiac signal processing chain.

It should be highlighted that we integrate many strategies for enhancing signal quality and robustness into the pipeline. These successive steps are essential for long-term monitoring in the

NICU. In this context, infants undergo various ventilation supports they need, frequent changes in body position, spontaneous movements and routine care, which could result in numerous and heterogeneous sources of artifacts and interferences as well as a particularly low Signal-to-noise ratio (SNR). They are even more critical in a population of premature infants that have very complicated while low cardiovascular variability with respect to adults.

The following sections describe each block of the workflow.

2.2.1 Signal quality evaluation and channel selection

In the specific case of real-life data, signal pre-processing of raw ECG recordings turns out to be necessary. The signals monitored in the protocol are continuous and inevitably noisy, and monitoring may last from days to months, depending on the hospitalization of the subjects in question. Consequently, during routine monitoring, the raw ECG waveforms may be alternately recorded among different ECG leads as presented in Section 2.1.3; besides, the raw signals are with various artifacts. Based on this, we divided the treatment of a raw ECG signal segment includes two steps, the first is to evaluate the signal quality of all the channels, and the second step is based on the quality indices to select the good channel with valid recording. In some segments, the signal quality of all channels may be unacceptable; there may also be some segments that happen to cover the moment of channel switching, so both channels contain valid data (as depicted in Figure 2.5). Ultimately, only the segments with signal quality exceeding a predetermined threshold can be passed to the next processing stage, which ensures data quality from the beginning.

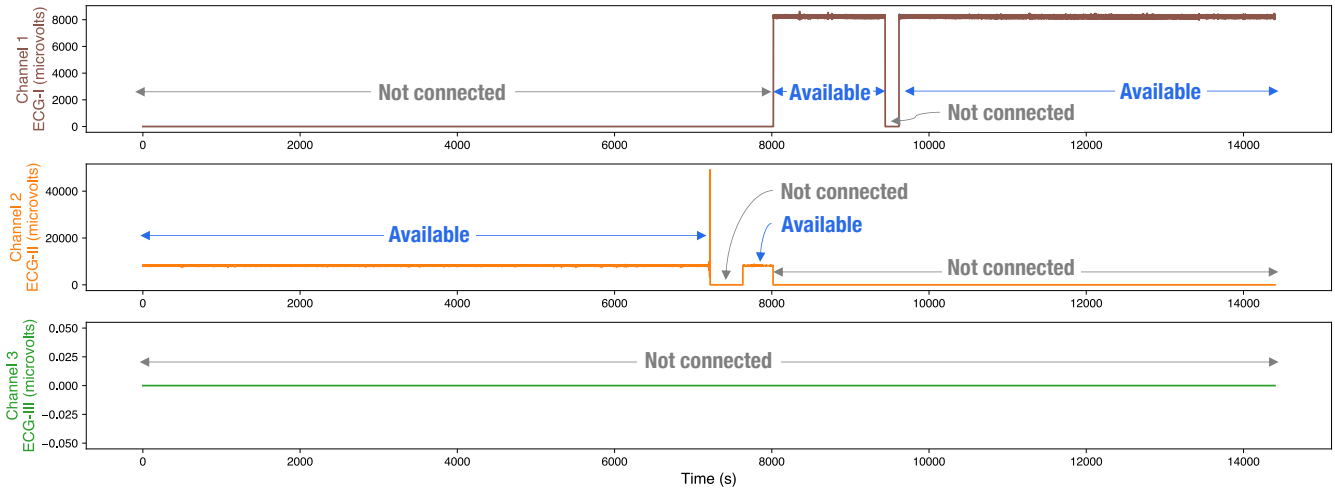

Figure 2.5: Example of a 4-hour 3-lead ECG recording with channel switching.

Noise detection

By observing the raw signals, we classify the noises into three types: impulses, saturation edges and flat segments. We design a simple yet effective time-domain signal processing approach

based mainly on sample derivatives (first and second order) and median filters. It is known that the derivatives represent the rate of change of the signal, and are often used to detect abrupt changes (impulses and edges) and enhance edges and fine details. The median filter is a classic filter used for signal smoothing and high-frequency noise reduction by replacing each value in the window with the median value. It is particularly effective at detecting impulse noise, which appears as sudden spikes or drops in the signal, and its feature to preserve the edges and details of the signals is quite useful in our case.

To detect and locate the noises, a set of threshold-based methods are proposed for detecting each type of noise. For clarity, we first define variables:

- *d1*: the first derivatives of the signal;
- *d2*: the second derivatives of the signal;
- *med_d1*: the median-filtered version of the first derivatives of the signal;
- *minD*: the minimum duration (in samples) of a segment considered as a flat segment, default is 250.

The detection strategy, with specific parameters for preterm infants, is as follows:

- Impulse detection: $Impulse \leftarrow abs(d2 \geq 5)$.
  Large changes in the *d2* exceeding a threshold of 5 is considered an impulse. If multiple impulses are detected close together, the segments between them are also marked as noises.
- Edge detection: $Edge \leftarrow abs(d1 \geq 5)$ & $abs(med\_d1) \leq 0.005$.
  Noises classified as edges are detected where the first derivative (*d2*) is large, but the median-filtered derivative (*med_d1*) is small. This indicates a sharp change followed by a flat segment, typical of saturation or disconnection.
- Flat detection: $Flat \leftarrow abs(d1 = 0)$ & $abs(med\_d1) = 0$, when flat duration $\geq minD$.
  Flat segments are identified where both the *d1* and *med_d1* are zero, only if the flat segments last longer than *minD*. There are segments where the signals show no variation, indicating possible disconnection, lead failure, or other issues where the ECG signal is not being properly recorded. The parameter *minD* helps to keep some short yet non-problematic flat segments that could occur naturally in the signal of extremely preterm newborns sampled at 500 Hz and after the derivatives and median filtering.

Note that there may be overlapping sections of the detected noises due to some morphological similarities among the three types of noise. Figure 2.6 shows an example of the noise location and removal, where the edge-type of noise is infrequently marked since it overlaps with the impulses sometimes. Furthermore, all the noise detection criteria are set for millivolts, so a unit conversion should be performed first if this is not the case.

After detection, we calculate the number of noisy samples and the percentages of each type of the three noises relative to the total length of the original signal segment. At the same time, the

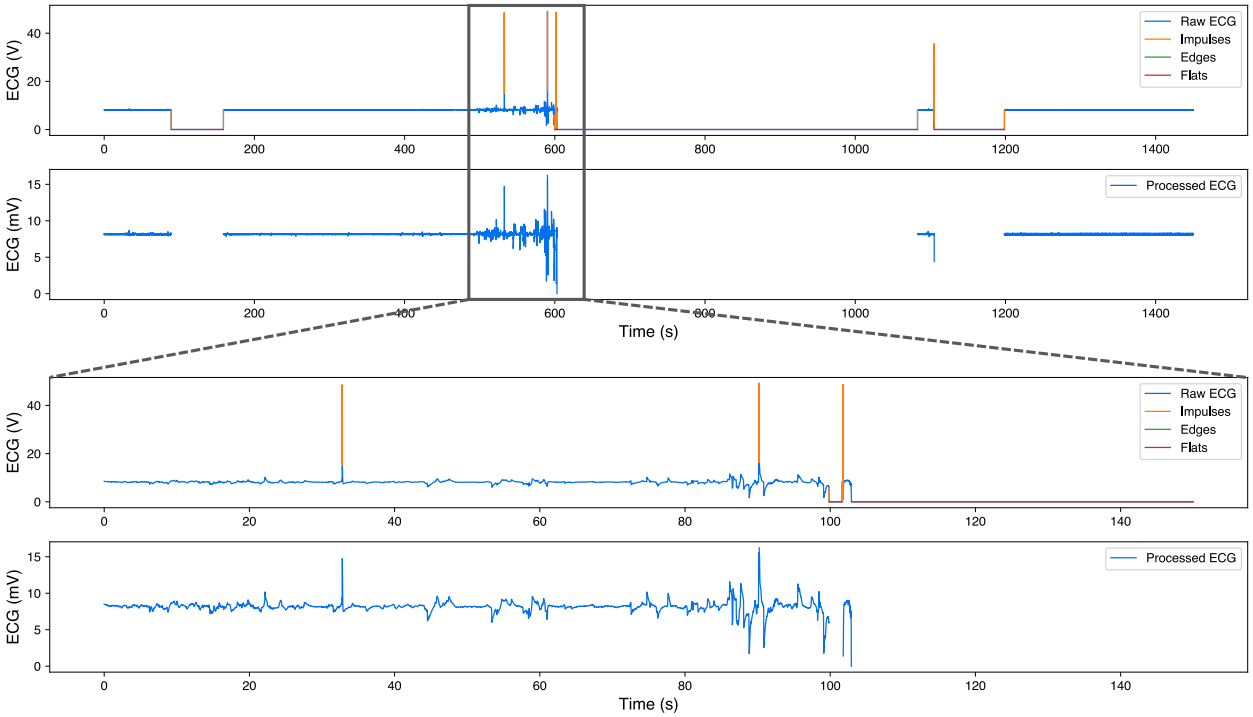

Figure 2.6: Illustration of the noise detection on an ECG segment with three types of noises.

noisy segments in the original signal are removed by replacing them with Not-a-Number (NaN) values.

Signal variability assessment - selection of the best channel

The signal quality indicators derived from the previous step are utilized for assessing the signal variability in all channels and thus choosing the best lead of ECG for further analysis, we develop three indices based on the results of noise detection.

Firstly, the index of the percentage of flat segments is used to determine whether a channel is "empty". Typically, during a certain duration of monitoring, at most one channel is connected to the infant and thus effectively records the ECG signal, while the other two channels are "empty" channels containing flat and zero-amplitude waveforms. Therefore, we first exclude the "empty" channels with a very high value of flat percentages, normally it will be 100%.

Then, we assess signal Signal-to-noise ratio (SNR) in each channel using an index based on the number of NaN values, a combination of the noises detected using the method described above. This index calculates the percentage of NaN values in the processed signal segment relative to the original signal length. A higher index value denotes a low SNR, which is undesired.

A third index measures how much a signal deviated from its baseline over time by calculating

the nested rolling standard deviation of a Z-score normalized signal, and we use it to quantify the signal variability. It first calculates the means and standard deviations within all windows with a predetermined length of 30 seconds and normalizes the signals in all the windows by Z-score normalization. Then the standard deviation of the Z-score normalized signal is calculated. Lastly, the final standard deviation, derived from the rolling standard deviation of the Z-score normalized signal, is used for evaluating the variability, or stability, of a signal. A low standard deviation indicates that the signal is relatively stable, while a high standard deviation suggests that the signal is more volatile and less stable, as compared in Figure 2.7.

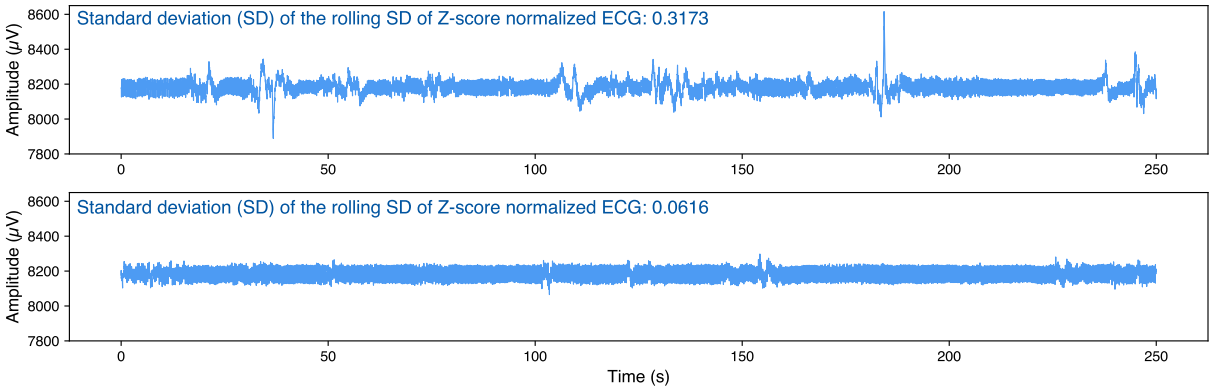

Figure 2.7: Comparison of the ECG variability with the nested rolling standard deviation.

The second and third measures are designed to choose the optimal one when more than one channel contains waveforms, i.e., a segment with channel switching. Together, these three indices served as criteria for selecting the best channel from a 3-lead ECG recording, i.e., the channel with the lowest flat percentage, NaN percentage, and standard deviation of rolling Z-score. Additionally, some limiting thresholds are set to reject a very poor quality signal, even if it is the "best" of the three leads, or to reject cases where all channels are "empty", which may happen when the patient is not connected to any of the electrodes corresponding to the leads.

2.2.2 QRS detection and RR interval extraction

Detecting the QRS complex is typically the first analysis in ECG signal processing, serving as the basis for identifying cardiac cycles and estimating cardiac markers such as heart rate, or for enabling further ECG segmentation and analysis. Figure 2.8 shows an illustration of normal cardiac cycles recorded in an ECG signal. The QRS complex is the most prominent deflection in the ECG signal, representing the electrical depolarization of the ventricles. The upward deviation of the QRS complex is the R wave, and the duration between two adjacent R waves is defined as the RR intervals (RRI). The heart rate (HR), the number of heart pulsations per minute, can be calculated from an ECG signal by dividing 60 by the duration of the RRI. This frequency is faster in preterm babies than in adults, due to the physiological evolution of the heart during the first weeks of life.

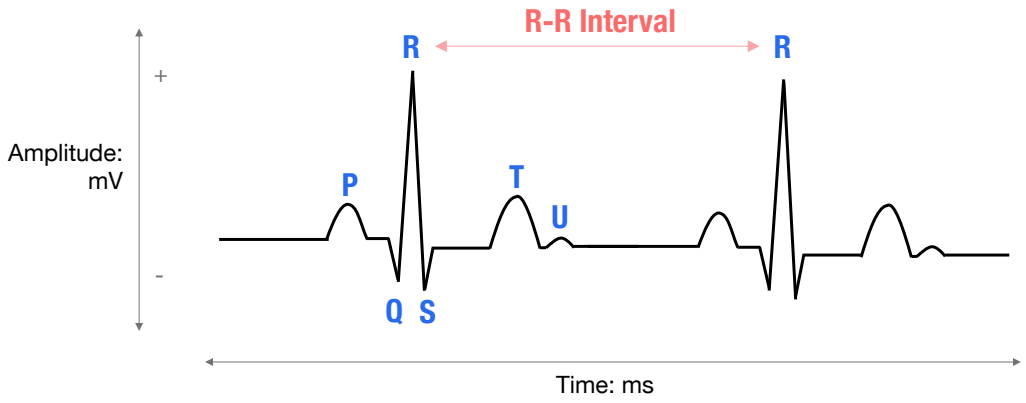

Figure 2.8: Illustration of normal cardiac cycles and the QRS complex.

Numerous methods have been proposed for detecting QRS complexes over several decades. The main proposed approaches include filtering and non-linear transformations [1], hybrid fuzzyneural networks [2], Stockwell transformatoion [3], wavelet analysis [4–7] and convolutional neural networks [8, 9]. Among those, our team previously proposed and evaluated a QRS detector implemented with a robust multi-feature probabilistic method, which consists of extracting a set of parameters and validating a detection by Bayesian filtering. This Multi-Feature Probabilistic Detector (MFPD) was adapted to the specific characteristics of the newborn electrophysiology, previously proposed and evaluated by our team [10].

We thus employ this MFPD QRS detector for the studies in the dissertation. Before being input into the detector, the selected ECG signals are first resampled to 1,000 Hz to match the frequency at which the detector parameters were optimized. The detector then detects and locates the R wave in the signal, and the RR interval series is obtained by calculating the time difference between two adjacent R waves. Figure 2.9 illustrates how to convert the detected R waves from an ECG signal into associated RR series. And Figure 2.10 shows a complete workflow from an ECG segment to the QRS detection and finally the extraction of RR interval series.

For clarification, we use the term "RR interval sequence" to refer to a numerically indexed sequence of RRI (e.g., 1,2,3⋯), and the term "RR or RRI series" to refer to a series of RR intervals that are indexed in time but not evenly spaced. Furthermore, when we talk about an "RR signal", it is a resampled version of the RR series and is, therefore, a uniformly spaced time series.

2.2.3 RR series correction

Limited by the SNR of the raw ECG signal and/or the accuracy of the QRS detector, the obtained RR interval sequences may inevitably have some anomalies or errors, which are well-known challenges affecting the HRV parameters. We develop a multi-step approach using logic rules based on pathological and rhythmic corrections to automatically correct (interpolating based

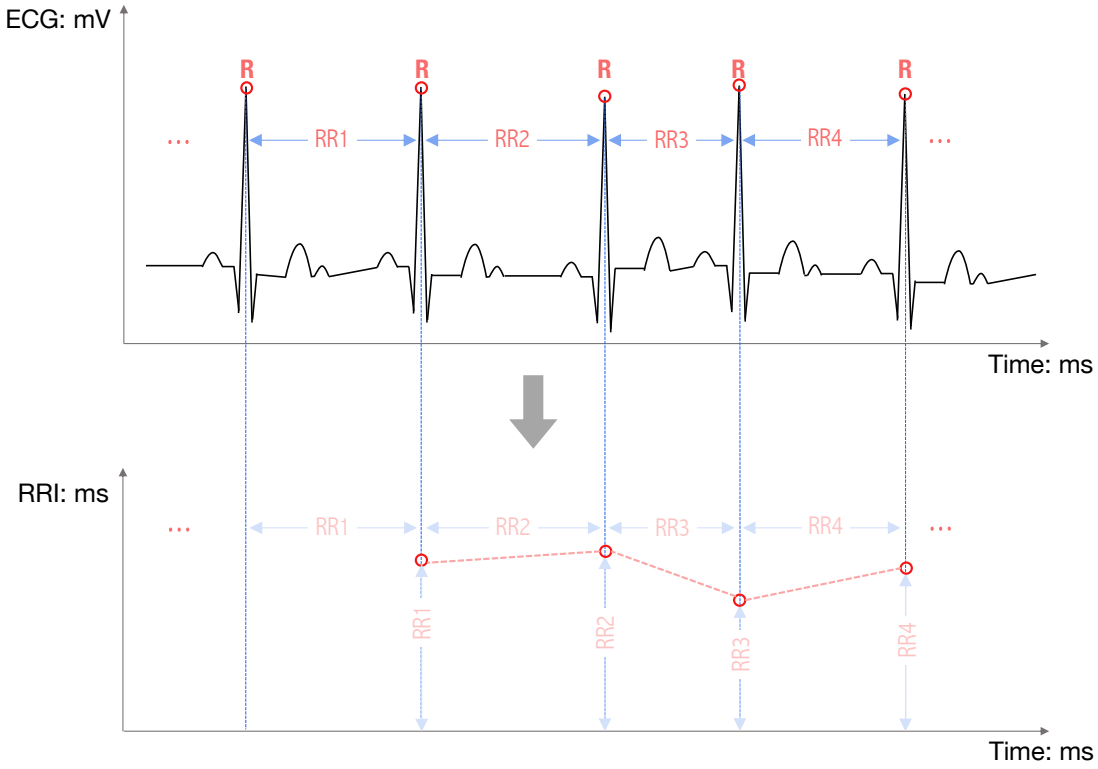

Figure 2.9: Illustration of the conversion from detected R waves to RR interval series.

on neighboring intervals) and/or reject (removing from the original RR sequences) artifacts and errors (including false positive and false negative detection of R waves).

Before defining different RR interval patterns, we first create a sliding-window view of the raw RR series with a size of 5 samples, this is an appropriate span to observe possible mistakes in the central R wave detection. Then a median filter is applied to the 5-sample segments of RR sequences in order to obtain a **baseline** trend in the given RR series. We also calculate two different sums on the sliding windows: **sumA** (the sum of the second and third RR intervals) and **sumB** (the sum of the second, third and fourth intervals). These three sets of values combined with the empirical soft thresholds are used as constraints to help determine which pattern the detection error belongs to. Different patterns have different identification rules and correction methods.

Correction patterns and rules

We classify various artifacts and erroneous detection patterns in R-wave detection into 10 categories, which, although we do not claim to enumerate all patterns, have been systematically expanded to include the major patterns that usually occur. Logic rules with specific thresholds corresponding to each of the patterns are developed to identify the detection error and correct it

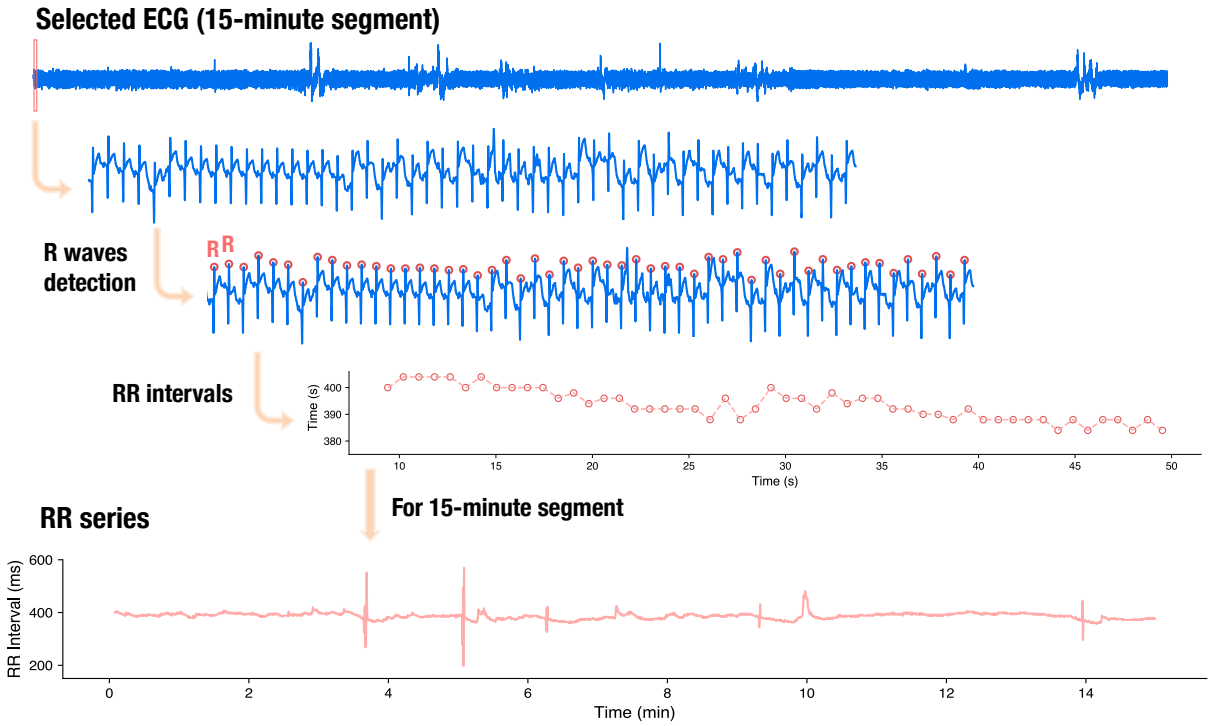

Figure 2.10: Workflow of ECG signal processing and RR interval series extraction.

by either replacing or interpolating. Besides, an intuitive scheme of each pattern is presented in order to demonstrate how it looks in an ECG signal, in which a "|" represents an R wave, a "!" is for false detection of the R wave, a "." is for missed detection of the R wave and "__" stands for a normal length of one cardiac cycle.

1. **Pattern_anyHigh**: any extremely high value (RR interval).
   Scheme: __|__|__.....__|__|__|__ (more than 10 missed R waves)
   When one RR interval's value is greater than 10 times the median RR interval, we consider it as a bad detection and simply replace it with the global median RR interval in the given RR series.
2. **Pattern_J0**: when missing too many R waves.
   Scheme: __|__|___....many....___|__|__|__ (more than 1 missed R wave)
   If the QRS detector misses multiple consecutive R waves, this is reflected in the ECG signal as a very long distance between two adjacent detected R waves (as illustrated in the scheme), which results in a very high RR interval value. A set of logic rules is designed to identify such long RR intervals, and the average of the neighboring RR intervals within a 5-sample window centered around the detected erroneous RR interval is used to replace this false detection.
3. **Pattern_J**: when missing several R waves, it is considered a valid detection.

Scheme: `__|__|_...several..._|__|__|__` (more than 1 missed R wave)
This pattern includes six subcategories recognizing the number of successive missed R waves ranging from 2 to 7. We divide the large RR interval by the number of missingness and use the derived average value to interpolate the original RR interval, by doing this, the overall length of the RR signal (in time rather than in index) remains unchanged. use the average We average the identified large RR intervals that fit this pattern by dividing them by the number of missed values and use the average to interpolate on the original RR intervals. By doing this, the overall length of the RR signal (in time, not in index) remains unchanged, which avoids the phase problem.

4. **Pattern_K**: 1 false detection + 1 missed detection.
   Scheme: `__|_!_.__|__|__|__` (when the mean is around the **baseline**)
   This pattern defines the cases as a combination of one false detection and one missed detection of R waves. It can only be classified into this category when the mean value of these two intervals, a short/long interval derived from the false detection and a long/short one due to the missed detection, is around the **baseline** RR interval that was initially calculated. Naturally, this pattern is corrected by replacing them with the mean value.
5. **Pattern_L**: (1 false detection + 1 missed detection) $\times$ 2.
   Scheme: `__|_!_._!_.__|__|__` (when the mean is around the **baseline**)
   Similar to the previous pattern, this pattern identifies one false detection followed by one missed detection, but this combination is repeated twice. The sum of the two false RR intervals and the following long interval (due to the missed R waves), i.e., the **sumB**, is divided by three and used to interpolate in between.
6. **Pattern_M**: 1 false detection + 1 missed detection + 1 false detection.
   Scheme: `__|_!_._!___|__|__` (when the **sumB** is around twice the **baseline**)
   One missed detection between two false detections of R waves is also a common pattern, resulting in three short intervals. In this case, the sum of these three detected RR intervals, i.e., the **sumB**, should be approximately twice the **baseline** interval. We then correct this pattern by replacing the first two intervals with half the abovementioned **sumB** and setting the last one to NaN value, which ensures the phase alignment.
7. **Pattern_N**: 1 missed detection + 1 false detection + 1 missed detection.
   Scheme: `__|__._!_._!__|__|__` (when the **sumA** is around three times the **baseline**)
   If there is only one R wave is detected during a long duration, this offers two relatively longer RR intervals, and if the sum (**sumA**) is as long as approximately thrice the **baseline**, then it can be classified as this pattern. To correct it, one-third of **sumA** is repeated three times in between two good intervals at the beginning and end of this pattern, as a replacement of the two longer intervals caused by this pattern.
8. **Pattern_O**: 1 false detection.
   Scheme: `__|_!_|__|__` (when the **sumA** is around the **baseline**)
   This is a simple pattern where one R wave is wrongly detected between two normal R waves,

leading to two short intervals. The sum of the short intervals (**sumA**) in this case should be around the **baseline** interval. Thus, we correct the first short interval caused by the false detection with the value of **sumA** and set another short interval as NaN value.

9. **Pattern_Q0**: consequent segments with many missed detection.
   Scheme: `__|_...many..._|_...many..._|_...many..._|__`
   When encountering consequent missed detection of many R waves (more than 6 or 8 R waves), very long RR intervals are derived. When the R waves fail to be detected and this occurs consecutively, many long RR intervals will occur. This pattern is an extension of **Pattern_J0** when the long missingness happens two or three times. We calculate the average value of two good RR intervals (one before and one after the long missingness) and directly replace the abnormally long intervals with the average. In this kind of poor detection, we treat it with simple replacement without considering the phase alignment. So the corrected RR series can be shorter than the original RR series if this pattern of detection exists.
10. **Pattern_Q**: consequent segments with several missed detection.
    Scheme: `__|_.several._|_.several._|_.several._|__`
    This pattern includes similar cases as in **Pattern_Q0** but with less missingness, which is feasible to be interpolated. It includes a total of 27 detailed subcategories recognizing different combinations of the number of successive missed R waves ranging from 1 to 5 and the number of such missingness episodes (2 or 3). Once the number of missed detections is identified, use the number to divide the corresponding long intervals and then interpolate the raw RR series with the average intervals.

Correction procedure

By using the rules and patterns listed above as building blocks, we develop a universal procedure to correct the raw RR series, as shown in Algorithm 1. It should be noticed that, instead of using the RRI series, we take the original output of the QRS detection, i.e., the list of indexes where R waves are located on an ECG signal (*R_index_on_ECG*), as the input of the correction procedure. The repeated **Pattern_J0** and **Pattern_J** in the end is for correcting the abnormal segments that are covered by the other patterns, for instance, consequent long RRI would contaminate the **baseline** which hindered the detection of the following long RRs. Besides, the step-by-step correction is performed twice, the first time is to correct the raw *R_index_on_ECG* (forward correction) and the second round takes the corrected series as a starting point, but in a backward manner.

To better demonstrate the efficiency of the proposed correction methods, we take an instance of raw RR series that was exported from a Philips Data Warehouse Connect included in a real-time monitoring system (this system will be introduced in Chapter 6). As it is real-life, uncleaned data, this RR series possesses almost all the patterns listed above, both the forward correction (upper panel) and backward (lower panel) correction are shown in Figure 2.11. It is quite clear that, from a zoom-out point of view, major artifacts are successfully rejected throughout the procedure.

Procedure 1 Procedure of RR correction.

Input: Raw *R_index_on_ECG*
Output: Corrected *R_index_on_ECG*

Temp ← *R_index_on_ECG*, Count = 2, *isReverse* = *False*
for Count > 0 **do**
  if *isReverse* **then**
    Temp ← *Temp* in reversed order
  else
    Do nothing
  end if
  Calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_J0**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_anyHigh**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_J**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_K**
  Apply **Pattern_L**
  Apply **Pattern_M**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_N**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_O**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_Q0**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_Q**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_J0**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Apply **Pattern_J**; Update *Temp*, calculate **baseline**, **sumA**, **sumB**
  Count = Count − 1, *isReverse* = *True*
end for
Corrected *R_index_on_ECG* ← *Temp*
return Corrected *R_index_on_ECG*

Nevertheless, there are some caveats in the rule-based correction approach. We correct false positive (wrong) and related false negative (missed) detection due to artifacts but not ectopic R waves due to arrhythmia or quasi-continuous jitters. Importantly, there are undoubtedly many more erroneous patterns that we have not yet included, depending on the population, patient care setting, database, etc. Thus, this correction is not a perfect approach, but it is still useful when detecting and thus accordingly correcting the most common artifacts and errors.

2.2.4 Stationarity analysis

Changes over time in the mean and variance of the corrected RR series are estimated and signal stationarity is tested. It is conducted by *i)* dividing the available RR series into shorter subsegments of 20 intervals, *ii)* calculating the mean on each subsegment, and *iii)* determining the variance of these mean values [11]. A low variance suggests the time series is relatively stationary. This stationary index can be used to select the best segment in terms of stationarity and lack of artifacts.

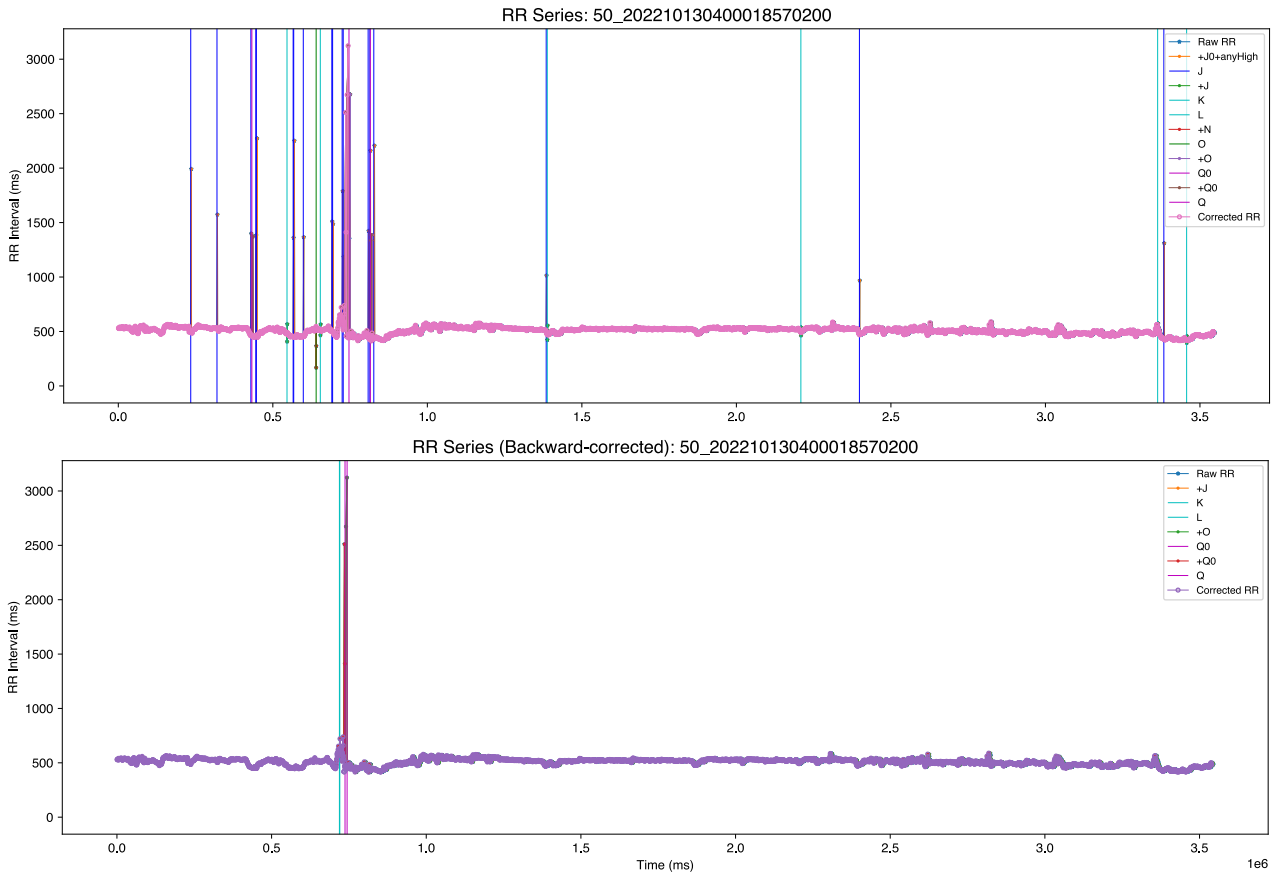

Figure 2.11: Example of proposed RR correction approach applied on a noisy RR series. Upper panel: the forward correction from raw RR series to a corrected result (pink). Lower panel: the backward correction takes the reversed forward-corrected RR series as input and performs a second round of correction, the corrected RR series is shown in purple.

2.2.5 HRV analysis

The instantaneous frequency of cardiac electrical and contractile activity, i.e., instantaneous heart rate, is largely regulated by the Autonomic nervous system (ANS). This neuro-modulation leads to Heart rate variability (HRV), defined by the fluctuation in the time intervals between consecutive heartbeats, i.e., RR intervals [12]. Therefore, HRV analysis is an interesting tool for studying cardiac function and the state of the autonomic nervous system in general, particularly the balance between sympathetic and parasympathetic (vagus) activities. In general, it is believed that an increase in sympathetic tone leads to a decrease in HRV, while an increase in parasympathetic activity leads to an increase in HRV. An optimal level of HRV is associated with health and self-regulatory capacity, and adaptability or resilience [13]. The analysis of HRV can be classified into three categories: time-domain, frequency-domain and non-linear metrics.

Time-domain parameters

Time-domain indices quantify the temporal patterns of HRV observed during monitoring periods.

- **Mean (ms)**: Mean RR interval in a given RR interval series.
- **Median (ms)**: Median RR interval in a given RR interval series.
- **Std (ms)**: Standard deviation of RR intervals in a given RR series.

$$\text{Std} = \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(RR_i - \overline{RR})^2} \tag{2.1}$$

where N is the number of RR intervals, RR_i is the i-th RR interval, $\overline{RR}$ is the mean of the RR intervals.
It is a global index of HRV reflecting all the long-term components responsible for the variability in the recording period.

- **Skewness (ms)**: It measures the asymmetry of the distribution of RR intervals.

$$\text{Skewness} = \frac{1}{(N-1)} \cdot \frac{1}{\sigma^3} \cdot \sum_{i=1}^{N}(RR_i - \overline{RR})^3 \tag{2.2}$$

- **Kurtosis (ms)**: It is the standardized fourth central moment of the samples and measures the "tailedness" of the distribution of RR intervals.

$$\text{Kurtosis} = \frac{1}{(N-1)} \cdot \frac{1}{\sigma^4} \cdot \sum_{i=1}^{N}(RR_i - \overline{RR})^4 \tag{2.3}$$

- **IDR (ms)**: Interdecile range of a given RR interval series. It measures the statistical dispersion, representing the range between the 90^{th} and 10^{th} percentiles of the RR intervals.
- **Rmssd (ms)**: Root mean square of successive RR interval differences.

$$\text{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2} \tag{2.4}$$

where N is the number of RR intervals, RR_i is the i-th RR interval.
It corresponds to the beat-to-beat variance in HR and is the primary time-domain measure used to estimate the parasympathetic nerves mediated changes reflected in HRV [14, 15].

- **pDec (%)**: Percentage of decelerated RR intervals. It is defined as the percentage of RR intervals larger than the mean RR interval of the past 50 successive intervals and this feature aims at explicitly extracting variations in HRV arising from decelerations [16].
- **stdDec (ms)**: Standard deviation of the decelerated RR intervals that contribute to pDec. It captures the magnitude of decelerations [16].
- **SAA (n.u.)**: Sample asymmetry analysis. It captures asymmetry in the histogram of RR

intervals and allows separate quantification of the contribution of accelerations and decelerations.

$$\text{SAA} = \frac{\beta \cdot \overline{(RR_i - med) \cdot \mathbb{I}(RR_i \geq med)}}{\alpha \cdot \overline{(med - RR_i) \cdot \mathbb{I}(RR_i < med)}} \tag{2.5}$$

where RR_i represents the individual RR intervals, med is the median of the RR intervals, α and β are weighting of deviations for HR accelerations (RR intervals less than the median) and decelerations (greater than or equal to the median), respectively, the default value is 2, $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the condition inside is true, and 0 otherwise, $\overline{(\cdot)}$ represents the mean value over the intervals [17].

- **AC (ms)**: Acceleration capacity. It is calculated as the average acceleration response with a segment length of 20 RR samples based on the phase-rectified signal averaging (PRSA) technique that detects quasi-periodic oscillations in time series data [18].
- **DC (ms)**: Deceleration capacity. It is the average deceleration response derived from the PRSA method. While DC is considered to reflect parasympathetic control of the sinus rhythm, the meaning of AC remains unclear [18].

Among these, the Rmssd, SAA, AC and DC are derived from the differences between consecutive RR intervals, and thus they primarily index high-frequency HR oscillations, remaining largely unaffected by long-term trends [13]. The SAA, pDec and stdDec are measurements specifically designed to capture neonatal HRV. Other indices such as the Triangular index (integral of the density of the RR interval histogram divided by its height) and pNN50 (percentage of successive RR intervals that differ by more than 50 ms) are not included in this dissertation. Indeed, these indices are rarely represented in neonatal literature and do not provide additional information compared to the main indices described above.

Frequency-domain parameters

Frequency-domain values determine the absolute or relative signal energy within specific frequency bands, which characterize the sympathovagal influences on the heart rhythm [19]. Different from the recommended spectral components for adults [19], newborns exhibit HRV fluctuations over a wide frequency range [20, 21]:

- High-frequency band between 0.2 and 2 Hz;
- Low-frequency band between 0.02 and 0.2 Hz;
- Very-low-frequency band between 0 and 0.02 Hz.

To better estimate the power spectral density (PSD) of the RR interval series, we first implement an interpolation on the RR series that creates a uniformly spaced time series sampled at 1,000 Hz, then down-sample the RR signal to 4 Hz. We use an autoregressive (AR) model with the Burg method to estimate the AR coefficients of the resampled RR signal by minimizing forward and backward prediction errors based on the reflection coefficient approach and Levinson recursion,

with an AR model order of 16 [22, 23]. Note that the mean of the RR signal is subtracted, as a simple form of detrending, before the estimation. The frequency analysis is implemented in Python using the *Spectrum* library [24]. After obtaining the PSD estimation, we calculate five frequency-domain parameters in the frequency bands given above:

- **LF (ms^2)**: Power spectral density in the low-frequency band.
  The LF component is modulated by both the sympathetic and parasympathetic nervous systems. Its interpretation has been more controversial and reflects combined changes in sympathetic and parasympathetic activity.
- **LFnu (n.u.)**: Normalized power spectral density in the low-frequency band.

$$LF_{nu} = \frac{LF}{LF + HF} \tag{2.6}$$

  The normalized value allows direct comparison of the frequency-domain measurements of different subjects despite wide variation in specific band power and total power among healthy, age-matched individuals [25].
- **HF (ms^2)**: Power spectral density in the high-frequency band.
  The HF component reflects parasympathetic activity and is generally called the respiratory band because it corresponds to the HR variations related to the respiratory cycle.
- **HFnu (n.u.)**: Power spectral density in the high-frequency band.

$$HF_{nu} = \frac{HF}{LF + HF} \tag{2.7}$$

- **LFHF (n.u.)**: Ratio of low-frequency power (LF) to the high-frequency power (HF).

$$LFHF = \frac{LF}{HF} \tag{2.8}$$

  It assesses the ratio between sympathetic and parasympathetic nervous systems activity [15]. A low LFHF ratio reflects parasympathetic dominance while a high ratio indicates sympathetic dominance [13].

Non-linear parameters

The non-linear analysis allows us to measure the complexity and unpredictability of the given RR interval series using mathematical techniques derived from chaos theory, fractal approaches, or information theory, etc.

Poincaré plot analysis of an RR series is the representation of all successive RR intervals as a function of the preceding RR intervals. This method can be described as follows: two adjacent RR intervals represent a point on the plot, the first RR interval (RR_i) represents the x coordinate, and the second RR interval RR_{i+1} represents the y coordinate [26]. By fitting all the points on the Poincaré plot into an ellipse, we can obtain some quantitative parameters from it [13].

- **SD1 (ms)**: Standard deviation of the distance of each point from the $y = x$ axis, specifying the width of the ellipse.
  It measures short-term HRV and correlates with high-frequency power. Mathematically, SD1 is identical to Rmssd [27].
- **SD2 (ms)**: Standard deviation of the distance of each point from the $y = x + \overline{RR_i}$, specifying the length of the ellipse.
  It measures both the short- and long-term HRV and correlates with low-frequency power.

Sample entropy, a modified version of approximate entropy, is a less biased and more reliable measure of signal regularity and complexity by quantifying the self-similarity of a time series [28, 29]. It is calculated as the negative natural logarithm of the conditional probability that two sequences similar to each other for m points, within a tolerance (r), remain similar when the sequence length is increased to $m+1$.

- **SampEn (n.u.)**: Sample entropy.

$$\text{SampEn}(m, r, N) = -\ln\left(\frac{A^{m+1}(r)}{B^m(r)}\right) \tag{2.9}$$

  where N is the length of the RR series, m is the embedding dimension indicating the length of sequences to be compared, r is the tolerance level (usually a fraction of the standard deviation of the given series), B is the number of pairs of sequences of length m that are within a distance r of each other and A is the number of pairs of sequences of length $m+1$ that are within a distance r of each other.

A higher SampEn indicates great complexity or high irregularity, while a lower SampEn value indicates more self-similarity and predictability in the series, commonly associated with disease. In the case of neonates, the parameters have been optimized to $m = 3$ and $r = 0.25$ [30].

Detrended fluctuation analysis (DFA) extracts the correlations between successive RR intervals over different time scales for measuring the fractal properties of the given series [31]. The process involves dividing the RR interval series into segments, detrending each segment, and calculating the root-mean-square fluctuation of the integrated and detrended time series. This is done over various segment lengths to determine the scaling exponent α. A Gaussian white noise (completely uncorrelated values) will yield an $\alpha \approx 0.5$, in contrast, a Brownian noise signal (strongly correlated series) will approach $\alpha \approx 1.5$, and series with $\alpha \approx 1$ display more fractal-like dynamics ($1/f$) [31]. We evaluated the fractal scaling exponent from two distinct ranges [25, 32, 33]:

- **α1 (n.u.)**: Short-range scaling exponent evaluated from 4 to 40 heartbeats.
- **α2 (n.u.)**: Long-range scaling exponent evaluated from 40 to 1000 heartbeats.

α1 and α1 provide different scopes into the balance between sympathetic and parasympathetic activity in the ANS. The short-range correlations extracted using DFA reflect the baroreceptor reflex, while long-term correlations reflect the regulatory mechanisms that limit fluctuation of the beat cy-

cle. A healthy ANS exhibits complex, fractal-like characteristics in RR fluctuations [34], reflecting adaptability to internal and external stressors.

2.3 Statistical Techniques

2.3.1 Correlation analysis - parallel relationship

Correlation measures the degree of association between two variables, indicating the extent to which the variables change concurrently, and usually, linearly. Note that this association is parallel, thus it does not necessarily imply a cause-and-effect relationship. The variables exhibit simultaneous variation without a distinction between independent and dependent variables, i.e., it is not about predicting Y from X.

Correlation coefficients, often denoted r or ρ, together with a p-value, quantify the degree and direction of correlation. The coefficient is a number ranging in $[-1, +1]$, where the sign indicates the direction of a relationship (negative or positive), and the larger the absolute value is, the stronger the correlation between the variables, and a correlation of 0 represents no linear relationship between variables, i.e., irrelevant. And the corresponding p-value provides information on whether the conclusions drawn from the data are meaningful, i.e., the correlation coefficient can be interpreted only if the p-value is significant [35].

Pearson correlation coefficient (Pearson's r)

The Pearson product-moment correlation coefficient [36], or Pearson correlation coefficient, is the most familiar measure of association. It is obtained by taking the ratio of the covariance of the two variables in question, normalized to the square root of their variances. Mathematically, it can be expressed as a division between the covariance of the two variables and the product of their standard deviations.

$$r_{X,Y} = corr(X, Y) = \frac{covX, Y}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[\,XY\,] - \mathbb{E}[\,X\,]\mathbb{E}[\,Y\,]}{\sqrt{\mathbb{E}\,[\,X^2\,] - (\mathbb{E}[\,X\,])^2} \cdot \sqrt{\mathbb{E}\,[\,Y^2\,] - (\mathbb{E}[\,Y\,])^2}} \tag{2.10}$$

where $\mathbb{E}$ is the expectation.

Spearman's rank correlation coefficient (Spearman's ρ)

Spearman rank correlation coefficient [37] is a non-parametric measure of rank correlation, where rank refers to the relative position label of the observations within the variable: 1^{st}, 2^{nd}, 3^{rd}, etc. It assesses how well two variables are monotonically related, i.e., as one variable increases, the other variable tends to increase, even if their relationship is not linear. Spearman's coefficient is appropriate for both continuous and discrete ordinal variables. Moreover, Spearman's correlation coefficient is a more robust method than Pearson's, i.e., more sensitive to non-linear relationships

[38]. The calculation of Spearmans's ρ follows the definition of Pearson's r (Equation 2.10) but it uses ranked variables rather than the raw data.

Shown in Figure 2.12 are some examples of scatter plots with different values of correlation coefficients, and both Pearson's r and Spearman's ρ are annotated for comparison. All results are calculated using the *SciPy* [39] library in Python.

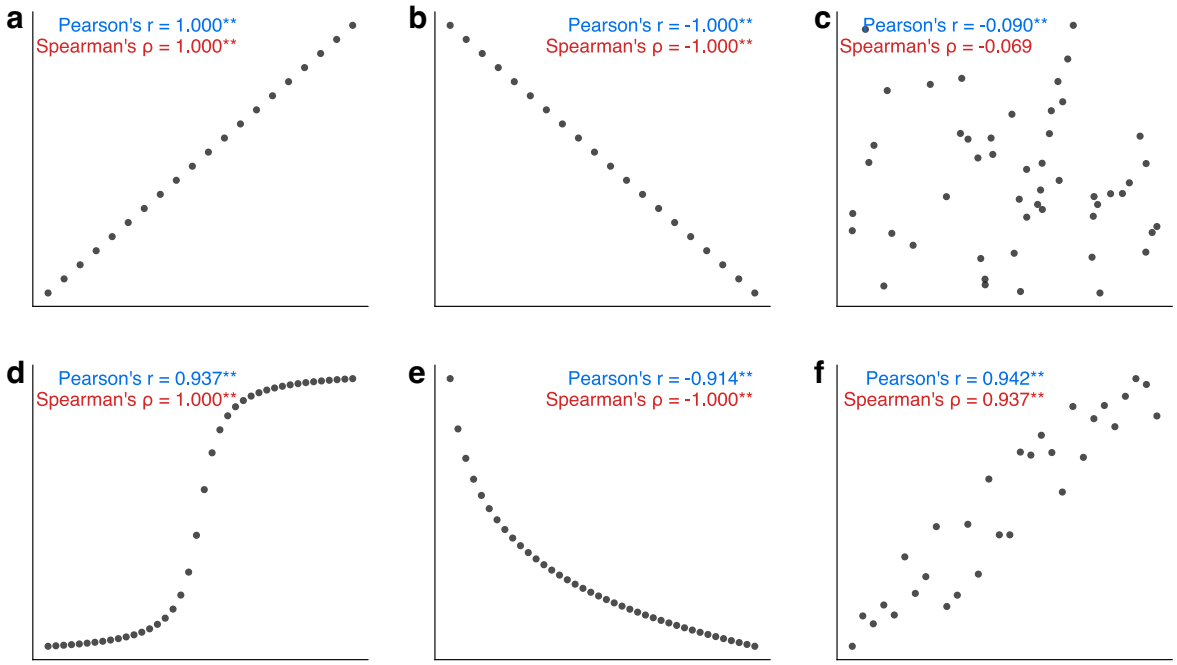

Figure 2.12: Examples of scatter plots with different correlation coefficients (r and ρ). (**a**) A linear increase. (**b**) A linear decrease. (**c**) A Gaussian white noise. (**d**) An arc-tangent function. (**e**) A logarithmic decrease. (**f**) A linear increase with some white noise. ** p-value $<<$ 0.01.

2.3.2 Regression analysis - dependent relationship

Regression analysis is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables. The primary goal of regression is to predict the dependent variable based on the values of the independent variables, while also understanding the strength and nature of these relationships. It should be noted that it is paramount to check whether there is a significant correlation between the variables before performing any linear regression, as correlation is a prerequisite of statistically meaningful regression results, that is, there is no regression without correlation.

A general formulation of a regression model, which proposes that $\boldsymbol{Y}_i$ is a function (regression

function) of $\boldsymbol{X}_i$ and β with e_i representing an additive error term, can be written as:

$$\boldsymbol{Y}_i = f(\boldsymbol{X}_i; \boldsymbol{\beta}) + e_i \tag{2.11}$$

The goal is to estimate the function $f(\boldsymbol{X}_i; \boldsymbol{\beta})$ that most closely fits the data. In order to conduct a regression analysis, the form of the function $f(\cdot)$ must be first specified, and different forms of tools are thereby used to estimate the parameters β, such as least squares [40], which estimates parameters $\boldsymbol{\beta}$ that minimize the sum of squared residuals.

$$\sum_{i=1}^{n} \left(f(\boldsymbol{X}_i; \boldsymbol{\beta}) - y_i\right)^2 \rightarrow \min_{\boldsymbol{\beta}} \tag{2.12}$$

It is also important to note that there must be adequate data to estimate a regression model in order to obtain a determined and only solution.

Regression models/functions can be broadly categorized into linear and non-linear types. Linear regression assumes that the relationship between the variables is linear, while non-linear regression is used when this relationship is more complex and cannot be adequately captured by a straight line.

Linear regression

Linear regression is one of the simplest and most commonly used types of regression, in which one finds the line (or a more complex linear combination, but need not be linear in the independent variables) that most closely fits the data according to a specific mathematical criterion. given a data set $\{y_i, x_{i1}, ..., x_{ik}\}_{i=1}^{n}$ of n random variables, a linear regression models the relationship between the dependent variable y and the vector of independent variables $\boldsymbol{x}$ using a linear equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon, \; i = 1, ..., n \tag{2.13}$$

where $\boldsymbol{x}_i$ $(x_{i1}, ..., x_{ik})$ are independent variables, $\boldsymbol{\beta}$ $(\beta_0, \beta_1, ..., \beta_k)$ are linear coefficients to be estimated and ε is the error term.

Linear regression models are often fitted using the Ordinary least squares (OLS) by the principle of least squares: minimizing the sum of the squares of the discrepancies or residual, $r_i = y_i - \hat{y}_i$, between the observed dependent variable (y_i) and the output of the linear function of the independent variables $(\hat{y}_i = \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}})$. OLS seeks to find the coefficients $\boldsymbol{\beta}$ that minimize the sum of squared residuals (SSR):

$$\mathrm{SSR} = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2 \rightarrow \min_{\boldsymbol{\beta}} \tag{2.14}$$

The simplest example of linear regression is to model a linear regression for n two-dimensional

sample points with only one independent variable x_i and one dependent variable y_i:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, ..., n \tag{2.15}$$

This is a straight line in a Cartesian coordinate system, in which the intercept (β_0) is such that the line passes through the center of mass ($\overline{x}$, $\overline{y}$) of the data points and the slope (β_1) indicates the degree of dependence of y_i on x_i.

Non-linear regression

Non-linear regression, on the other hand, is used when the model function has a non-linear combination in the parameters. The model, $f(\cdot)$ in Equation 2.11, can take various forms, such as exponential, logarithmic, polynomial, or more complex functional forms. The goal, similar to linear regression, is to estimate the parameters such that the model fits the data as closely as possible. The least squares analysis can also be used to solve non-linear regression problems by non-linear least squares. By choosing initial values for the parameters, the algorithm refines the parameter iteratively, and at each iteration, the system is approximated by a linear one, and thus the core calculation is similar in both linear and non-linear cases. Since the parameters are obtained by successive approximation, and thus non-linear regression generally does not have a closed-form solution.

Robust non-linear least square method

In reality, we are coping with data from real life, and data is thus often contaminated by outliers or extreme measurements. In this case, the least-squares solution can become significantly biased to avoid very high residuals on outliers.

One of the well-known robust estimators is the L1-estimator, which minimizes the sum of the absolute values of the residuals. The only disadvantage of the L1-estimator is the function is non-differentiable everywhere, which is particularly troublesome for efficient non-linear optimization.

Based on this, researchers [41, 42] proposed an approach incorporating differentiable features and robustness in the estimation process. They introduced a scalar-valued, sub-linear function $\rho(\cdot)$, where "sub-linear" means its growth should be slower than linear, in order to reduce the influence of outlier residuals and contribute to the robustness of the solution, and the function $\rho(\cdot)$ is referred as a loss function.

There are several options for the loss function, we can choose according to the specific problem to be solved for the degree of robustness requirements. Here we list three representative functions:

- Linear function which gives a standard least squares: $\rho(z) = z$
- Smooth approximation to absolute value loss, "soft L1 loss": $\rho(z) = 2(\sqrt{1+z} - 1)$

This function is relatively mild and gives approximately absolute value loss for large residuals.

- Cauchy loss: $\rho(z) = ln(1+z)$
  This one is strongly sub-linear and gives significant attenuation for outliers.

Furthermore, the value of the soft margin used to differentiate between inliers (significant data points) and outliers (less significant or noisy data points) is 1.0 by default, which is an unweighted threshold. To generalize, a scaling parameter C for the loss function $\rho(\cdot)$ was introduced and evaluated as:

$$\hat{\rho}(r_i^2) = C^2\rho\left(\left(\frac{r_i}{C}\right)^2\right) \tag{2.16}$$

C scales the residuals r_i before passing them into the $\rho(\cdot)$ function, the larger the C, the smaller the adjusted residuals $\frac{r_i}{C}$, leading to a less sensitivity the loss function is to outliers (larger residuals). When residuals are small, $\hat{\rho}(r_i^2) \approx \rho(r_i^2) \approx r_i^2$ holds for any $\rho(z)$ defined above. Therefore, in robust regression, C serves as a parameter that controls the trade-off between sensitivity to outliers and the overall fit. By adjusting C, it can influence the degree to which the loss function penalizes larger residuals (potential outliers), affecting the robustness of the regression model.

Overall, a robustified bound-constrained non-linear least-squares optimization problem can be formulated as:

$$\begin{aligned} &\sum_{i=1}^{n}\rho(r_i^2) = \sum_{i=1}^{n}\rho\left(\left(f(\boldsymbol{x_i};\boldsymbol{\beta}) - y_i\right)^2\right) \rightarrow \min_{\boldsymbol{\beta}} \\ &\text{Subject to} \quad \text{lb} \leq \boldsymbol{\beta} \leq \text{ub} \end{aligned} \tag{2.17}$$

where r_i is the residuals, $\rho(\cdot)$ the loss function controlled by scaling parameter C, lb and ub are the lower and upper bounds for the estimated parameters, respectively. The bounds are set to ensure the estimated parameters are within reasonable ranges and sometimes help converge more quickly to the optimal solution.

In brief, robust non-linear regression is designed to fit a non-linear model that describes the majority of data, and the robustness is improved by giving the data different weights thereby reducing the influence of outliers on the solution, through the loss function $\rho(z)$. The whole idea of robust non-linear least squares is implemented in the *SciPy* [39] Python library.

2.3.3 Consistency analysis - agreements and differences

Consistency analysis is a statistical method used to evaluate the agreement or reliability between different sets of measurements, observations, or ratings. Consistency analysis functions in three major aspects:

1. Comparison of methods:
   To compare different methods of measurement or analysis to ensure they produce similar results. This can be critical when validating a new measurement technique against a gold

standard.

2. Reliability and reproducibility:
   To ensure that measurements or evaluations are reliable, reproducible, and can be trusted across different conditions or evaluators. Without consistency, conclusions drawn from data may be questionable.
3. Inter-rater and intra-rater reliability:
   To assess how consistently different raters (inter-rater reliability) or the same rater at different times (intra-rater reliability) score the same observations in studies involving human evaluators (e.g., clinicians, judges).

Several methods are developed to conduct consistency analysis, such as Bland-Altman analysis, Intraclass correlation coefficient (ICC) and Cohen's kappa, etc. We will focus on the Bland-Altam analysis in the following section.

Bland-Altman analysis

The Bland-Altman (B&A) analysis is a method to assess the agreement between two different instruments or two measurement techniques [35, 43, 44]. This approach helps identify any systematic bias between the methods and assesses the consistency of the measurements across their range.

Given two measurement methods M_1 and M_2, the B&A analysis involves calculating several statistics:

- **Difference:** the differences between the results of M_1 and M_2;
- **Mean:** the averages of the results of M_1 and M_2;
- **Mean difference (bias):** the mean of the differences, representing the systematic bias between the two methods, e.g., one method systematically obtains higher or lower observations than the other;
- **Limits of agreement (LoA):** the LoA are typically set at **Mean difference** $\pm 1.96\times$ standard deviation (SD) of the **Difference**.
  These limits indicate the range within which 95% of the differences between the two methods are expected to fall. These limits provide a threshold for acceptable agreement: if most data points fall within these limits, the two methods are considered to agree sufficiently.

Graphically, the B&A plot was introduced by Bland and Altman [43] to visualize the differences against the means as a scatter plot, which is the most common way to plot, providing a clear and intuitive representation of the agreement for easier comparison. The y-axis of such a plot shows the difference between the two paired measurements ($M_1 - M_2$) and the x-axis represents the average of these measures ($(M_1 + M_2)/2$). The horizontal dashed lines indicate the ± 1.96 SD.

We illustrate this with example data points as shown in Figure 2.13a. The bias (mean difference) of -27.17 units is presented by the gap between the horizontal solid line and the zero

differences (y-axis is zero). This negative bias seems caused by the measurements over 200 units, while for lower observations, the data derived from the two methods are closer to each other (scatters are around zero differences when the mean is lower than 200 units). The agreement limits are from −94.24 to 39.91 units.

It is also possible to plot the differences as percentages or ratios, and one can use the first method or the second one, instead of the mean of both methods. The option of plotting the differences as percentages is useful when there is an increase in variability of the differences as the magnitude of the measurement increases. As shown in Figure 2.13b, for our example samples, the bias (mean difference) is −17.40%, almost constant for all the measured concentrations, with the exception of very low values. The agreement limits are from −91.9% to 57.1%.

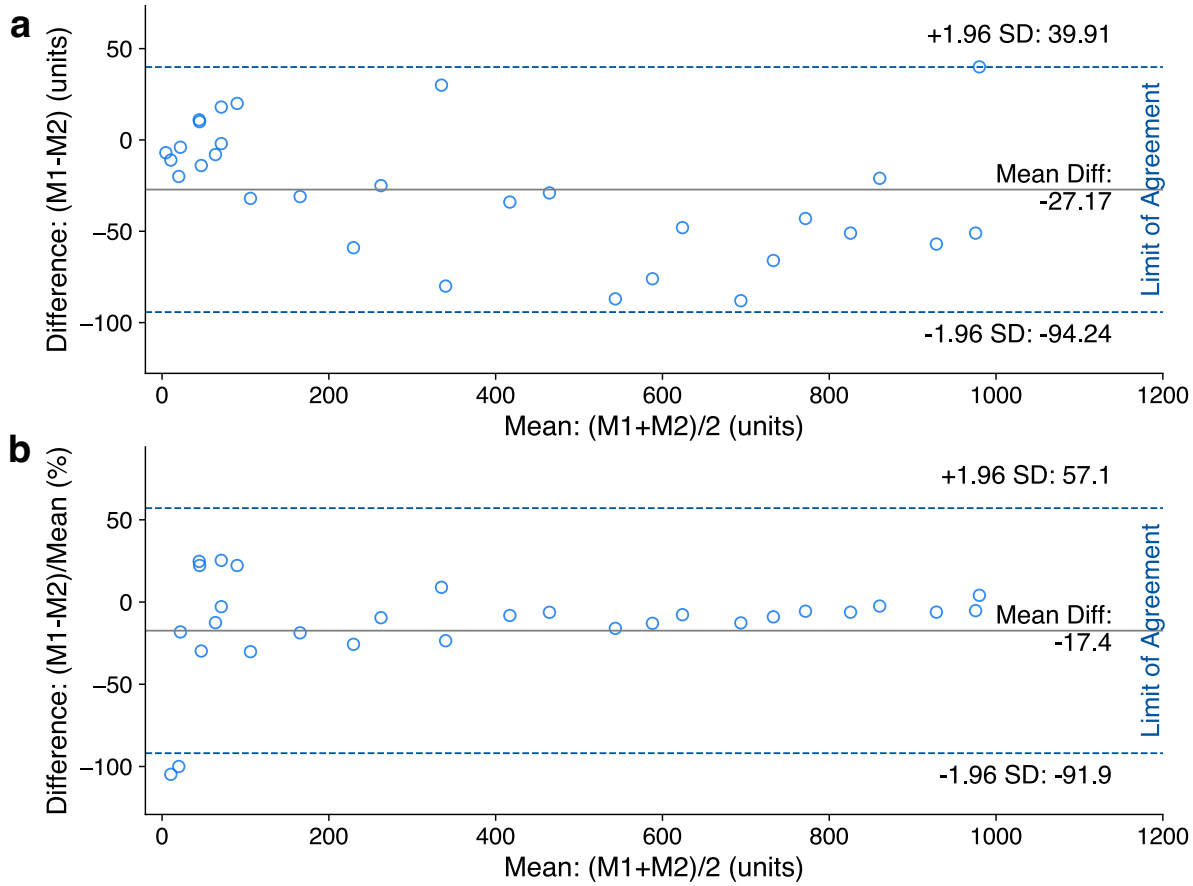

Figure 2.13: Illustration of Bland and Altman plots with representation of bias and limits of agreement (dotted line), from −1.96 SD to +1.96 SD. (**a**) B&A plot of unit values: differences between method M_1 and M_2 against the mean of the two measurements. The bias is −27.17 units and the LoA is 67.07 units. (**b**) Relative B&A plot of percentage values: differences expressed as percentages between method M_1 and M_2 against the mean of the two measurements. The bias is −17.40% and the LoA is 67.07 units.

The B&A plot simply quantifies the bias and a range of agreement, within which 95% of the differences between one measurement and the other are included. It does not indicate whether the agreement is sufficient or suitable for using one method (M_1) or the other (M_2) indifferently. To interpret the results, i.e., whether the agreement is too wide or sufficiently narrow, depends on a priori the limits of maximum acceptable differences (limits of agreement expected), based on specific studies and purposes that provide biologically and analytically relevant criteria. We use the B&A analysis to obtain the statistics to see if these limits are exceeded the expected LoA, or not.

Apart from the information plots themselves, we can also integrate correlation analysis (Section 2.3.1) and linear regression analysis (Section 2.3.2) into the consistency analysis. For example, fitting a regression line of the differences on the B&A plots could help in detecting a proportional difference [45].

2.3.4 Outlier detection

Identifying outliers in data is a critical step in many statistical analyses and research endeavors, as outliers can significantly influence the results and interpretations of a study. Outliers can distort parameter estimates in statistical models, leading to biased results, which is also a justification for the significance of robustness, as we discussed in Section 2.3.2. In data science, whether to develop statistical models or machine learning models, identifying and handling outliers is a always crucial step in data cleaning, before performing any analysis.

In the dissertation, we test different approaches to detect outliers from the studied dataset. If considering the nature of the variable involved, the outlier detection methods can be classified into four branches: univariate methods, bivariate methods, multivariate methods and mixed or specialized methods.

Univariate methods

Univariate methods refer to methods that involve only one variable or one feature from a dataset without considering relationships with other variables. Basic statistical methods such as Z-score and Interquartile range (IQR) are often used for detecting outliers univariately.

The Z-score measures how many standard deviations a data point is from the mean of the dataset, for a data point x, the Z-score is

$$Z = \frac{x - \mu}{\sigma} \tag{2.18}$$

where μ is the mean and σ is the standard deviation of the data. Typically, data points with a Z-score greater than 3 or less than -3 are considered outliers. This is an easy way but with the assumption that the data follows a normal distribution, which is quite strict, especially for small datasets.

The IQR measures the statistical dispersion in a dataset and is calculated as the difference between the first quartile ($Q1$, 25th percentile) and the third quartile ($Q3$, 75th percentile): $IQR = Q3 - Q1$. Outliers are defined as data points that fall below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$. This method does not assume a particular distribution and is more robust to skewed data than the Z-score method.

The IQR is also included in a box plot (or box-and-whisker plot), which is a graphical representation of data. A box plot visualizes the distribution of a dataset by displaying its minimum, $Q1$, median (second quartile, $Q2$), $Q3$ and maximum. The whiskers in the box plot extend to the smallest and largest values within 1.5 times the IQR from $Q1$ and $Q3$, respectively. The same principle is used to identify outliers in the box plots as the IQR method, that is, all data points outside the whiskers are outliers. Both the IQR and box plot methods are non-parametric, which do not assume any specific data distribution.

Multivariate methods

As datasets become increasingly complex, often featuring multiple variables variables and high-dimensional features, the limitations of univariate outlier detection methods become apparent. While univariate approaches are useful for detecting outliers in single variables, they fall short of capturing the intricate relationships between variables in multivariate datasets. We discuss four/three methods to identify outliers in a multivariate manner.

PCA-based outlier detection

We propose a method on the basis of Principal component analysis (PCA) to identify outliers using a Gaussian distribution model. PCA is a dimensionality reduction technique that transforms a dataset with many variables into a smaller set of uncorrelated variables known as principal components (PC), which captures the most variance in the data [46]. By reducing the data to the first P principal components (PC_1,..., PC_P, P is normally much less than the original number of dimensions), this algorithm simplifies the multivariate data into a P-dimensional space. In this transformed new orthogonal space, the mean vector and the covariance matrix are calculated for the selected P principal components, which are used to model a multivariate normal distribution.

To detect outliers, we assume that the data points follow a multivariate normal distribution in this transformed space. The probability density function (PDF) of the multivariate normal distribution is evaluated for each data point in the transformed PCA space. The samples with a PDF value below a predefined threshold are identified as outliers.

This method capitalizes on the concentration of most data points within a certain range and identifies those that fall far outside this range as anomalies or outliers. However, it has some limitations as it is based on the assumption of normality and is somehow sensitive to threshold choice.

Mahalanobis-Chi-Square method

Unlike Euclidean distance, Mahalanobis distance accounts for correlations between variables. It effectively measures the distance of a data point from the mean of a distribution, taking into account the covariance among variables. It is defined as:

$$D^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \tag{2.19}$$

where x is the data point, μ is the mean vector of the distribution, Σ is the covariance matrix of the variables and Σ^{-1} is the reverse of the covariance matrix.

A Chi-square distribution with degrees of freedom equal to the number of variables (dimensions) in the dataset is then used to compare with the squared Mahalanobis distance (D^2) [47]. To identify outliers, critical values from the Chi-square distribution for a given significance level (α) and degrees of freedom serve as a set of thresholds. Data points with squared Mahalanobis distance (D^2) exceeding this critical value (thresholds) are considered outliers.

K-Means (normal clusters versus abnormal clusters)

K-Means clustering [48] is a for partitioning data into clusters based on similarity. It assigns data points to clusters such that the variance within each cluster is minimized. When setting the K as 2, one for inliers and another for outliers, K-Means can be used as a good approach to identifying abnormal samples. The idea is that the cluster that captures the majority of the data points is considered the inlier cluster, in which the samples are closely grouped around the cluster centroid, representing the implicit typical pattern of the dataset. A cluster that has very few data points or unusually high variance might be considered abnormal. Data points that belong to these clusters but are far from the centroid are likely to be outliers.

Specialized method

Angle-Based Outlier Detection (ABOD) [49] is a technique used to identify outliers in a dataset by leveraging the geometric properties between data points, particularly in high-dimensional spaces. ABOD detects outliers by examining the distribution of angles formed between a given data point and its neighbors. The variance of its weighted cosine scores to all neighbors can be considered as the outlying score.

In detail, for each data point in the dataset, two sets of vectors are first calculated: data point-to-neighbors vectors and neighbor-to-neighbor vectors for all pairs of neighbors. The identification of the neighbors could be performed using k-nearest neighbors algorithm. Then the associated angles can be derived by applying dot product and *arccosine* to the vectors. The variance of the angles is regarded as the outlying score and is used to measure the typical angle distribution around each data point. Accordingly, high variance indicates that the data point and its neighbors have a sig-

nificantly different angular distribution compared to the majority of the data points, suggesting that the point is an outlier. Finally, a threshold is used to classify points as outliers based on their ABOD scores. The threshold can be set based on the statistical properties of the score distribution or empirical validation.

ABOD is shown to be effective in high-dimensional data where traditional distance-based methods might struggle. And since it focuses on angles rather than absolute distances, it is also less sensitive to feature scaling.

As for handling the detected outliers, the most direct way is to exclude them from further analysis, this is an efficient option when it does not introduce any selection bias on the dataset but with a cost of losing potential valuable information. Other strategies for dealing with outliers, which can preserve the integrity of the datasets, include transforming data (e.g., scaling) to reduce the influence of outliers without removing them and others or using robust methods that are less sensitive to outliers, etc.

2.4 Machine Learning Algorithms

Machine learning (ML) is a branch of Artificial Intelligence (AI) [50] concerning the development and studying of statistical and non-statistical-based algorithms that can learn from data, identify patterns and generalize to new, unseen data, enabling them to make decisions *without explicit programming* [51] and facilitating automation and decision-making processes across various industries. During training, a learning algorithm iteratively adjusts the model's internal parameters to minimize errors in its predictions [52].

ML algorithms can be broadly categorized into the following types based on different learning paradigms based on the nature of the "signal" (*feature*) or "feedback" (*label*) provided to the learning system [53]:

- **Supervised Learning**
  The algorithm learns from labeled data, where the input-output pairs are given. The goal is to learn a general rule that maps inputs to outputs. Supervised learning has two primary objectives: one is to classify data points to predefined classes or categories using developed models (*classification*), and the other is to predict continuous outcomes based on input features (*regression*).
- **Unsupervised Learning**
  The algorithm works with unlabeled data and tries to identify the underlying patterns or structures in its inputs. Common tasks include clustering and dimensionality reduction.
- **Semi-supervised Learning**
  This approach combines both labeled and unlabeled data for training, leveraging a small amount of labeled data along with a large amount of unlabeled data. This type of algo-

rithm is useful in tasks such as protein sequence classification and web content classification, where labeling data is expensive or time-consuming.

- **Reinforcement Learning**
  The algorithm learns by interacting with a dynamic environment and performing actions by an agent, receiving feedback in the form of rewards or penalties, and aiming to determine the optimal actions to take in a specific context to maximize the cumulative reward over time. It is often used in subjects of robotics, autonomous vehicles and game AI (e.g., AlphaGo).

Each algorithm has its strengths and limitations and no single algorithm works for all problems [54], so the choice of algorithm depends on factors such as the nature of the data, the problem at hand, and the desired trade-off between interpretability and performance.

In clinical settings, ML techniques are increasingly being used to solve tricky and complex issues clinicians have posed. In the next sections, we give an overview of those that have been employed in this dissertation or that have most often been used in recent research aimed at improving neonatal care.

2.4.1 Supervised learning algorithms

Some common supervised learning algorithms are briefly introduced in this section, some are developed for only classification tasks or regression tasks whilst some are able to handle both.

Naïve Bayes (NB)

Naïve Bayes (NB) is a simple probabilistic classifier based on Bayes' Theorem, which assumes independence between features. The algorithm is called "naïve" because it makes the assumption that all features (or input variables) are independent of each other, which is rarely true in real-world data. Despite this strong and often unrealistic assumption, NB works effectively in tasks such as text classification and spam filtering [55].

The algorithm calculates the posterior probability of a class given the features by combining the prior probability of the class and the likelihood of observing the features under that class. It selects the class with the highest posterior probability as the predicted output. There are several variants of NB, depending on how the likelihood is computed, including Gaussian Naïve Bayes for continuous data, Multinomial Naïve Bayes for discrete or count data (commonly used in text classification), and Bernoulli Naïve Bayes, which is ideal for binary features.

Due to its simplicity, Naïve Bayes is highly scalable and works well with large datasets, although its performance can be limited when feature independence is not a reasonable assumption. On the flip side, although NB is known as a decent classifier, it is known to be a bad estimator [56].

Logistic regression (LR)

The Logistic Regression (LR) [57] is a linear classification model and is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. The LR estimates the probability that a given input belongs to a particular category by fitting data to a logistic function (also known as the sigmoid function) with regularization. This function maps any real-valued number into a value between 0 and 1, making it ideal for binary classification tasks and can also be extended to multi-class settings.

Decision tree (DT)

Decision Tree (DT) is a tree-like model where decisions are made by splitting the data into subsets based on feature values, forming a hierarchy of rules. Data scaling or normalization is not required in tree-based algorithms. In a Decision Tree, each internal node represents a feature (or attribute) on which the data is split, and each branch represents an outcome of the decision based on that feature. The terminal nodes, known as leaves, represent the final decision or output (either a class label in classification or a value in regression).

Commonly used decision tree algorithms include ID3, C4.5 and CART. These three algorithms use different methods to select the optimal splitting attributes, with the goal of ensuring that the samples in each branch node of the decision tree are as similar as possible in terms of class. This aims to increase the "purity" of each node as the tree is split further. ID3 [58] splits the nodes based on information gain (entropy) and handles only categorical data for classification tasks. As an extension of ID3, C4.5 [59] handles both categorical and continuous data and uses the gain ratio for splitting and supports pruning to prevent over-fitting. The CART [60] generates binary trees (i.e., each node has two children) regardless of the number of classes, making it simpler and more structured. CART suits for both classification (binary and multi-class) and regression tasks, and it uses Gini impurity (for classification) or variance reduction (for regression) to find the optimal splits.

Decision trees are easy to interpret and visualize, making them popular for understanding decision-making processes. However, they can be prone to over-fitting, especially with complex data, which is often mitigated by pruning or using ensemble techniques like random forest.

Random forest (RF)

Ensemble methods combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability and robustness over a single estimator. Two very representative ensemble methods are Random Forest (RF) (Bagging strategy) and eXtreme Gradient Boosting (XGBoost) (Boosting strategy).

Random Forest (RF) [61] is a powerful ensemble supervised machine learning algorithm,

which integrates multiple base learners, i.e., decision trees or Classification and negression tree (CART) in specific, to improve accuracy and robustness [60]. RF follows the same principle as in a single CART, using the Gini index as a measure of purity to control the growth and division of the trees.

The ensemble strategy of random forest adopts a parallel method: the Bagging [62]. Different sub-training sets are obtained from the entire training data through multiple sampling with replacement (bootstrap). Several base learners (CART) with "certain accuracy" are obtained through parallel training and remain independent. If it is a classification task, the output of the RF comes from the majority vote of each tree in the forest; if it is a regression task, the output is the average results of all ensemble decision trees.

Its advantages over a single decision tree and other ML algorithms involve the introduction of two types of randomness:

- **Random selection of training samples.**
  The training samples used for each decision tree in a random forest are selected using the bootstrap resampling method. While the samples chosen for training each tree are the same size as the original training set, they are not identical due to the randomness introduced by the bootstrap procedure. This randomness creates a certain degree of variance in the information learned by each tree, while still maintaining consistency and overlap among them, thereby promoting diversity within the model.
- **Random selection of training features.**
  In each decision tree (CART), only a random subset of features is selected from the original feature set for training. When splitting nodes, the algorithm considers only this randomly chosen subset to determine the optimal split. This randomness in feature selection helps reduce the correlation between individual trees, which enhances the overall generalization performance of the algorithm.

The introduction of these two forms of randomness effectively minimizes model variance. As a result, RFs achieve a high level of resistance to over-fitting and exhibit strong generalization capabilities without requiring additional pruning, and are well-suited for processing high-dimensional datasets. Also, due to the parallel generation of trees, the computational complexity of training a RF is on par with that of training a single DT model, underscoring its efficiency as an ensemble algorithm.

Many hyper-parameters of RF are shared across both classification and regression tasks, with some specific to each task. The key hyper-parameters of Random Forest includes:

— **Shared hyper-parameters:**

- **n_estimators**: The number of trees in a forest.
  Increasing this generally improves performance but also raises computational cost. However, the performance improvements will diminish after reaching a certain number of trees.

- **max_features**: The number of features to consider when splitting a node.
  This can be set to "sqrt", "log2", or a fraction, controlling the randomness in each tree. Lower values reduce variance but may increase bias.
- **max_depth**: The maximum depth of each tree.
  Limiting this parameter helps prevent over-fitting by controlling tree complexity.
- **min_samples_split**: The minimum number of samples required to split a node.
  Higher values limit tree growth and help reduce over-fitting.
- **min_samples_leaf**: The minimum number of samples required at a leaf node.
  This prevents trees from growing too deep and helps generalize better.

— **Classification-specific hyper-parameters:**

- **criterion**: The function used to measure the quality of a split.
  For classification, common options are "gini" (Gini impurity) or "entropy" (information gain), determining how splits are evaluated to improve class purity.
- **class_weight**: Weights associated with classes in imbalanced datasets.
  This parameter adjusts the model to handle imbalanced datasets by giving more importance to minority classes.

— **Regression-specific hyper-parameters:**

- **criterion**: The function used to measure the quality of a split.
  For regression, this is typical "mse" (mean squared error) or "mae" (mean absolute error), determining how splits are evaluated to reduce prediction errors.
- **min_impurity_decrease**: A node is split if the impurity decrease is greater than this value.
  This helps control the granularity of tree splits and makes split criteria more finely.

Extreme gradient boosting (XGBoost)

Different from the Bagging strategy for RF, the underlying foundation of eXtreme Gradient Boosting (XGBoost) [63] is the Boosting strategy, i.e., each base learner is generated in a sequential fashion with strong dependencies on each other. With multiple optimizations on the GBDT algorithm [64], its speed and efficiency have reached the extreme, making it a flexible and efficient XGBoost algorithm, which is also where the name "eXtreme Gradient Boosting" comes from.

XGBoost combines multiple weak learners (typically DTs) to create a strong learner. Each tree in the model is trained sequentially, and the objective is to correct the errors made by the previous trees by minimizing the residuals. XGBoost includes L1 (Lasso) and L2 (Ridge) regularization terms to prevent over-fitting, making it more robust than standard gradient boosting algorithms.

XGBoost is widely used in tasks like classification, regression and ranking due to its flexibility and performance, and is known for its performance, scalability, and ability to handle large datasets with high accuracy. It has become a preferred algorithm in tasks requiring high accuracy and computational efficiency, such as financial modeling, customer behavior prediction, and medical

diagnoses. It has been particularly successful in winning numerous machine learning competitions due to its versatility, speed, and precision.

Multilayer perception (MLP)

Multilayer perceptron (MLP) [65] is a type of Artificial neural networks (ANN) used for both classification and regression tasks. It is a fundamental deep learning model that consists of multiple layers of nodes, or "neurons", where each neuron mimics a biological neuron by applying an activation function to the weighted sum of inputs.

The MLP consists of an input layer, an output layer and one or more hidden layers in between. The input layer receives the input data, with one neuron for each feature. The hidden layers perform computations by applying weights and biases to the inputs, followed by activation functions. The activation functions, such as *ReLu*, *sigmoid* or *tanh*, introduce non-linearity to help capture complex patterns in the data. The model's learning capacity increases with the number of hidden layers and neurons. The output layer produces the final output. For classification, the output typically represents class probabilities, while for regression, it provides continuous values. The MLP features a feedforward architecture, where information flows in one direction—from the input layer, through hidden layers, to the output layer—without any cycles.

The training of MLP uses a backpropagation algorithm, which adjusts the model's weights by minimizing the difference between predicted and actual outputs, typically using gradient descent. This process allows MLPs to learn and refine their predictions over time. Additionally, MLPs have the ability to approximate any continuous function, thanks to their "universal approximation" property, making them versatile for a wide range of machine learning tasks.

Shallow convolutional neural networks (Shallow CNN)

Shallow Convolutional neural networks (CNN) are a type of neural network designed primarily for processing grid-like data, such as images. Unlike deep CNN, which consists of many layers of convolutional operations and complex architectures, shallow CNN have a simpler network structure and fewer network parameters. Some variants of shallow CNN have been proposed mainly targeting image classification tasks, including a shallow network that combines CNN with Support vector machine (SVM) [66], a shallow CNN with logarithmic filter groups [67], and a novel shallow CNN with only 4 layers with small size of convolution kernel to accelerate training convergence and improve the accuracy[68], etc.

The architecture of shallow CNN typically consists of a few convolutional layers followed by pooling layers and a final fully connected layer. The convolutional layers apply filters to extract features from the input data while pooling layers reduce dimensionality and retain important information. The fully connected layer interprets the features extracted by the convolutional layers

and makes the final predictions. For network training, shallow CNN are trained using backpropagation and optimization algorithms like stochastic gradient descent. They adjust weights to minimize the loss function, which measures the difference between predicted and actual values.

Shallow CNN are often used for tasks where the complexity of deep networks is not necessary, such as basic image classification, object detection in simpler contexts, and feature extraction in the preliminary stages of more complex pipelines. While shallow CNN may not perform as well as deeper networks on complex tasks, they provide a good starting point for understanding convolutional neural networks and can be effective for simpler applications. In fact, with the advantage of convolutional filters, shallow CNN can also efficiently capture local patterns and trends in time series data with task-specific convolutional kernels. In Chapter 5, where shallow CNN are employed, we will further describe the shallow CNN that are designed for sepsis early detection.

Mixed-effects random forest (MERF)

Mixed-Effects Random Forest (MERF) represents an advanced extension of traditional random forests, designed to handle complex hierarchical and multi-level data structures. This approach integrates the strengths of RF with mixed-effects models to account for both fixed and random effects in data. In supervised learning, MERFs are particularly valuable when dealing with data that exhibit inherent hierarchical or nested relationships, such as repeated measurements, clustered data, or data with multiple levels of variability, where complex dependencies and interactions have arisen. By incorporating both fixed effects, which capture systematic patterns across the entire dataset, and random effects, which account for variability at different levels (such as within clusters or groups), MERFs can provide more nuanced and accurate predictions.

The very initial origin of mixed-effects models is Linear mixed-effects model (LMM) [69], a statistical model that incorporates fixed and random effects to accurately represent non-independent data structures. LMM offers an alternative to traditional analysis of variance (ANOVA) that often assumes that observations within each group are independent. This assumption, however, may not hold in cases of non-independent data, such as nested, hierarchical, longitudinal, or clustered datasets. Non-independent data occurs when the variability between outcomes is influenced by correlations within or between groups. For instance, when studying healthcare methods involving multiple hospital systems, there are multiple levels of variables to consider, such as individual patients within different departments of hospitals and hospitals within larger health networks. In such cases, a solution to modeling hierarchical data is using Linear mixed-effects models, allowing us to understand the important effects between and within levels while incorporating the corrections for standard errors for non-independence embedded in the data structure [70, 71].

Using a similar idea of mixed effects but not limited to simple linear models or parametric statistical models, Hajjem et al. [72] adapted Regression Tree (RT) algorithms, i.e., CART, for clustered data with continuous outcomes using a mixed-effects approach. Their Mixed-Effects Re-

gression Tree (MERT) algorithm demonstrated significant improvements over standard trees in simulations, particularly when random effects are non-negligible. The key concept of MERT is to dissociate the fixed from the random effects. Essentially, it involves iterative calls to a standard RT algorithm within the Expectation-Maximization (EM) framework [73, 74]. In each iteration, a standard RT is constructed from transformed response data, with the current estimate of the random effect component subtracted from the original response. Independently, Sela and Simonoff [75] proposed a similar method, known as Random Effects Expectation-Maximization (RE–EM) trees.

One step further, Hajjem et al. proposed a generalization of RFs to clustered data consisting of replacing the RT within each iteration of their MERT algorithm with a forest of trees, and named this method as Mixed-Effects Random Forest (MERF) [72]. The MERF was defined as follows:

$$\begin{aligned} &y_i = f(X_i) + b_i Z_i + \epsilon_i, \\ &b_i \sim N(0, D) = N(0, \sigma_b^2),\ \epsilon_i \sim N(0, R_i) = N(0, \sigma_e^2 I_{m_i}),\ i = 1, ..., m \end{aligned} \tag{2.20}$$

where y_i is a $(m_i \times 1)$ vector of responses for the m_i observations in cluster i; X_i is the fixed-effects covariate matrix of size $(m_i \times p)$; Z_i is the random-effects covariate matrix of size $(m_i \times q)$, where q is the number of random effects; b_i is a scalar of $(q \times 1)$ representing the linear coefficients of random effects and it should be estimated for each cluster i and the covariance matrix of b_i is D; and ϵ_i is the $(m_i \times 1)$ vector of errors and its covariance matrix is R_i. Importantly, the non-linear function $f(\cdot)$ is estimated using a standard Random Forest using the information in X_i. The random part, $Z_i b_i$, is assumed linear. The total observations number is $N = \sum_{i=1}^{m} m_i$. Other assumptions in the MERF algorithm are that b_i and ϵ_i are independent and normally distributed and that the between cluster observations are independent.

The MERF is implemented under the framework of the maximum likelihood Expectation-Maximization (EM) algorithm, which was originally applied for LMM [69]. The main difference lies in the fact that the linear regression step used for the fixed effects is replaced by the application of the Random Forest algorithm. Briefly, the method consists of the application of the following steps:

1. Set $r = 0$. Let $\widehat{b}_{i(0)} = 0$, $\widehat{D}_0 = I_q$ and $\widehat{\sigma}^2_{e(0)} = 1$.
2. Set $r = r + 1$. Update $y^*_{i(r)}$, $\widehat{f}(X_i)_{(r)}$ and $\widehat{b}_{i(r)}$.
   (a) $y^*_{i(r)} = y_i - Z_i \widehat{b}_{i(r-1)}, i = 1, ..., m$;
   (b) Estimate $\widehat{f}(X_i)_{(r)}$ using the RF algorithm [61], with X_i as the inputs and $y^*_{i(r)}$ as the target;
   (c) Compute $\widehat{b}_{i(r)}$ as detailed in [76].
3. Update $\widehat{\sigma}^2_{e(r)}$ and $\widehat{D}_{(r)}$ analytically, as detailed in [76].
4. Keep iterating by repeating steps 2 and 3 until convergence.

The convergence of the algorithm is evaluated by computing at each iteration a Generalized log-likelihood (GLL) criterion [76].

To predict the response for a new observation j within a known cluster i used in fitting the MERF model, we combine its population-averaged RF prediction, $\widehat{f}(X_{ij})$, with the predicted random effect for that cluster, $Z_i\widehat{b}_i$. However, if the new observation belongs to a cluster not present in the original training data, the prediction relies solely on the population-averaged RF prediction, $\widehat{f}(X_{ij})$, since no cluster-specific random effects are available. Hence,

- For a known cluster: Prediction $= \widehat{f}(X_{ij}) + Z_i\widehat{b}_i$.
- For a unknown cluster: Prediction $= \widehat{f}(X_{ij})$.

2.4.2 Hyper-parameter optimization

Hyper-parameter optimization is a critical step in ML that involves selecting the best set of hyper-parameters for a model to improve its performance and generalization ability. Unlike model parameters, which are learned from the training data, hyper-parameters are set before the training process begins and influence how the learning algorithm performs. Depending on the algorithms, hyper-parameters are different, varying from the learning rate, the number of hidden layers in a neural network, or the number of trees in a random forest.

The most widely used method to conduct hyper-parameter optimization is grid search. This method involves specifying a range of values for each hyper-parameter (hyper-parameter space) and exhaustively evaluating all possible combinations (in the training data). Although thorough, it can be computationally expensive, especially with large datasets and numerous hyper-parameters.

Random search is a lighter alternative to grid search. Instead of testing all possible combinations, random search samples combinations randomly from specified hyper-parameter space. It is often more efficient than grid search and can yield comparable results with less computational cost.

Successive halving [77] is yet another method to find the optimal hyper-parameter combinations. It is an iterative selection process where all candidates (the parameter combinations) are evaluated with a small amount of resources at the first iteration. Only some of these candidates are selected for the next iteration, which will be allocated more resources. Only a subset of candidates that consistently rank among the top-scoring candidates across all iterations can "survive" until the last iteration. Each iteration is allocated an increasing amount of resources per candidate.

Bayesian optimization uses probabilistic models to guide the search for optimal hyper-parameters [78, 79]. It builds a model of the objective function and uses it to choose promising hyper-parameter values, balancing exploration and exploitation.

Evolutionary algorithms, inspired by natural evolution, can be used to optimize hyper-parameters through processes such as mutation, crossover, and selection [80]. They are particularly effective in complex search spaces and can adaptively search for suitable hyper-parameter settings.

In summary, effective hyper-parameter optimization can significantly enhance model performance by finding the best configurations for learning algorithms. It helps in reducing over-fitting or under-fitting and ensures that models are well-tuned to the data they are trained on.

2.4.3 Model performance evaluation

Model evaluation strategies

When evaluating machine learning models, it is crucial to ensure that the model generalizes well to new, unseen data. Several techniques are commonly used to estimate how well a model is likely to perform on future data. These include train-test split, cross-validation, and Monte Carlo cross-validation (also known as repeated random sub-sampling).

The train-test split is the most straightforward technique for evaluating a model. In this approach, the dataset is divided into two distinct sets: a training set and a testing set. Typically, 70%-80% of the data is used for training the model, and the remaining 20%-30% is used to test its performance. This method is quick and easy to implement but has limitations. Since only one subset of the data is used for testing, the results can be highly dependent on how the data is split. If the split is unrepresentative, the model's performance may be overestimated or underestimated. Moreover, in small datasets, this method can waste valuable data by leaving a large portion unused during training.

Cross-validation is a more robust technique than the train-test split. It involves splitting the dataset into multiple subsets (or folds) and repeatedly training and testing the model on different portions of the data. One of the most common forms is k-fold cross-validation. It splits the dataset into k equal-sized folds, trains the model on $(k-1)$ folds, tests it on the remaining fold, and repeats the process k times, averaging the results to provide a more reliable performance estimate. This method ensures that every data point is used for both training and testing, and it provides a more stable estimate of the model's performance. However, cross-validation can be computationally expensive, especially for large datasets or complex models. Besides, stratified k-fold cross-validation is a variant often used when the dataset is imbalanced. It ensures that each fold has approximately the same class distribution as the original dataset, making the evaluation more representative.

Monte Carlo cross-validation, also known as repeated random sub-sampling, is another method for evaluating models. Unlike k-fold cross-validation, Monte Carlo cross-validation randomly splits the dataset into training and testing sets multiple times, and for each split, the model is trained and evaluated. This process is repeated n times, and the model's performance is averaged across all repetitions. This approach provides flexibility in how the data is split (e.g., using different train-test ratios for each iteration) and allows for more repetitions compared to k-fold cross-validation. However, it may result in an overlap between training and testing sets in different iterations, which can lead to a slightly biased result.

Choosing between these methods depends on the dataset size, the model's complexity, and the resources available. Cross-validation is generally preferred for more reliable results, while Monte Carlo cross-validation provides flexibility and randomness, and the train-test split is used when computational resources or time is very limited.

In addition to model evaluation strategies, different metrics are used for quantifying the performance of the model itself. We divide the metrics according to the specific needs of different tasks and problems.

Classification evaluation metrics

For classification tasks, where the goal is to predict categorical outcomes, the results can be summarized in a confusion matrix. The confusion matrix is especially useful when the dataset has more than two classes as it clearly visualizes the distribution of predictions across multiple classes. Figure 2.14 illustrates a confusion matrix of a binary classification problem by providing a detailed breakdown of the model's correct and incorrect predictions across all classes. The true positive (TP) and true negative (TN) are the numbers of instances correctly classified as positive and negative, respectively; while the false positive (FP) indicates the actual negative instances incorrectly predicted as positive, also known as a "Type I error"; and false negative (FN) is the number of instances incorrectly predicted as negative, also known as a "Type II error".

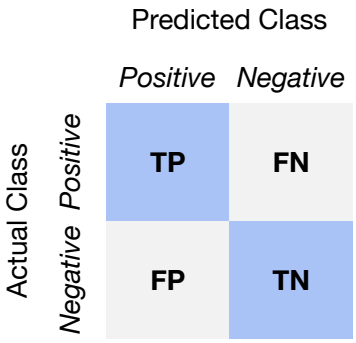

Figure 2.14: Confusion matrix for a binary classifier. TP: true positive. FP: false positive. FN: false negative. TN: true negative.

Several important performance metrics can be derived directly from the confusion matrix, including accuracy (Acc), sensitivity (or recall or True positive rate (TPR)), specificity (True negative rate (TNR)), False positive rate (FPR), precision (Positive predictive value (PPV)), F1-score and Balanced accuracy (BAcc):

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.21a}$$

$$\text{Sensitivity (Se, Recall)} = TPR = \frac{TP}{TP + FN} \tag{2.21b}$$

$$\text{Specificity (Sp)} = TNR = \frac{TN}{TN + FP} \tag{2.21c}$$

$$FPR = 1 - Sp = \frac{FP}{FP + TN} \tag{2.21d}$$

$$\text{Precision} = PPV = \frac{TP}{TP + FP} \tag{2.21e}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.21f}$$

$$\text{BAcc} = \frac{Se + Sp}{2} \tag{2.21g}$$

While useful for balanced datasets, the Accuracy can be misleading in imbalanced data scenarios, as it may favor the majority class. In contrast, other metrics mentioned above are more informative for imbalanced data. These metrics focus on the performance of the model in correctly identifying the minority class and managing trade-offs between false positives and false negatives, offering a more nuanced evaluation of model performance when one class significantly outnumbers the other.

In addition, measures derived from two curves are also commonly used as performance metrics. The first is the Area under the receiver operating characteristic curve (AUROC) and it is particularly useful when classes are balanced. The Receiver operating characteristic curve (ROC) itself plots the True positive rate against the False positive rate at various thresholds. The area under this curve (AUROC) is a single value that summarizes the performance of the classifier. The AUROC value ranges from 0 to 1: a value of 0.5 indicates a classifier that performs no better than random guessing; a value of 1 indicates a perfect classifier that separates classes with no errors. The closer the AUROC is to 1, the better the model is at distinguishing between the classes.

Another metric, Area under the precision-recall curve (AUPRC), which is more suitable for evaluating models on imbalanced datasets. The Precision-recall curve (PRC) plots precision on the y-axis and recall on the x-axis, showing how they change across different thresholds. The AUPRC captures the area under this curve, providing a summary of model performance. Unlike AUROC which has a baseline of 0.5 (random guess), the baseline of AUPRC is determined by the proportion of positive instances in the dataset. In other words, the baseline AUPRC corresponds to the precision achieved by a random classifier, which is equivalent to the prevalence of the positive class in the test data. Baseline AUPRC is typically low when the dataset is highly imbalanced (i.e., the positive class is rare). When comparing the performance, instead of directly comparing the absolute values of AUPRC, it should be compared against this baseline to assess how much better it performs than random guessing. The farther the AUPRC is above the baseline, the better the model is at distinguishing the positive class from the negative class.

Regression evaluation metrics

Regression models, which predict continuous outcomes, are evaluated using different metrics that focus on the accuracy of the predictions relative to the true values. Mean squared error (MSE) measures the average squared difference between predicted and actual values. It penalizes larger errors more significantly due to the squaring, making it sensitive to outliers. Root-mean-square error (RMSE) is the square root of MSE, giving error values in the same units as the target variable, which makes it easier to interpret. Mean absolute error (MAE) measures the average absolute difference between predicted and actual values. It gives equal weight to all errors, making it less sensitive to outliers compared to MSE. The R-squared (R^2) metric, also known as the coefficient of determination, indicates how well the model's predictions explain the variability of the target variable. It ranges from 0 to 1, where 1 indicates perfect predictions.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{2.22a}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{2.22b}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{2.22c}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{2.22d}$$

2.5 Interpretability and Explainability Analyses

As machine learning systems become increasingly sophisticated and their applications more critical, understanding how these models arrive at their predictions or decisions is essential for ensuring trust, accountability, and fairness.

EXplainable artificial intelligence (XAI) is a crucial area of research and practice in ML and AI that focuses on making complex models more interpretable and understandable.

Explainability refers to the ability of an ML or AI model to provide transparent and understandable insights into its decision-making process. The objective here is thus to make the internal workings of the model comprehensible to humans, so they can understand why the model made a particular prediction or recommendation. Explainability involves providing information about how the model arrived at its conclusions.

Model **interpretability** is the first step towards model explainability. Interpretability refers specifically to the ability of an ML or AI model to provide insights into its internal workings without necessarily providing a detailed explanation. In other words, interpretability focuses on un-

derstanding how the model works at a high level, rather than delving deep into the specifics of each decision. Interpretability is often achieved through two families of techniques:

- Model visualization: Visualizing the model's architecture or behavior to understand its overall structure and relationships. Attention maps are one common example of this technique.
- Feature importance analysis: Identifying which input features are most important for the model's predictions.

While interpretability provides a general understanding of how an AI system works, explainability goes further by providing detailed insights into specific decisions or outcomes. This section introduces various techniques and methodologies used to enhance the model interpretability.

2.5.1 Variable importance

Variable importance (VI) is a key concept in interpretability and explainability analyses, particularly in complex models such as machine learning algorithms, where understanding the contribution of each input variable to the model's predictions is essential. The goal of Variable importance analysis is to quantify the relative influence of each input feature on the model's output, providing insights into how the model makes decisions. Understanding which variables are most important can significantly enhance the transparency of the model and support interpretation. This is particularly valuable in fields such as healthcare, finance, and policy-making, where explainability is crucial for validating model outputs and ensuring trust. Moreover, identifying the most influential variables enables feature selection, model simplification, and improved generalizability by reducing model complexity without significant loss of predictive power.

In practice, variable importance can be measured in several ways depending on the type of model. For linear models, the magnitude and sign of the coefficients provide direct insight into the importance of each variable. A larger absolute value of a coefficient indicates a stronger influence on the target variable, with positive or negative signs indicating the direction of the effect. For example, in a linear regression model, a large positive coefficient means that an increase in the corresponding variable will strongly increase the predicted outcome, while a large negative coefficient suggests a strong inverse relationship. In contrast, for non-linear models such as decision trees, random forests, or neural networks, variable importance is typically assessed through different techniques, as these models do not rely on linear relationships. Methods like Gini importance (used in decision trees) or permutation importance (where the values of a feature are shuffled to assess the impact on model performance) are commonly applied. These techniques estimate how much each feature contributes to reducing prediction error or improving model accuracy, even when complex interactions between variables are present.

We use the Random Forest (RF) algorithm, which is primarily used in this dissertation, as a starting point to introduce two RF built-in variable importance estimation methods based on mean decreased Out-of-bag (OOB) accuracy and mean decreased Gini impurity. Moreover, we will also

briefly introduce permutation variable importance that provides a model-agnostic approach to estimate the features' contributions to the model performance.

Mean decreased out-of-bag accuracy

The importance of individual features can be determined by calculating the average decrease in the accuracy of Out-of-bag samples. The RF classifier or regressor is trained using bootstrap aggregation, where each new tree is fit from a bootstrap sample of the training observations, and this leads to a collection of Out-of-bag samples that were not used for model training. Considering the number of a bootstrap sample is N, the probability that a sample is not be sampled is $\left(1-\frac{1}{N}\right)^N$, and when $N \to \infty$, we have:

$$\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^N \approx \frac{1}{e} \approx 0.368 \tag{2.23}$$

Thus, the OOB samples comprise approximately one-third of the overall training samples. The OOB error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RF to be fit and validated whilst being trained [81]. The OOB error is often considered an unbiased estimate of the random forest generalization error, which approximates the k-fold cross-validation [61] that requires extensive computation, and it can be used to compute the importance of individual feature variables. The importance of a feature in an Random Forest (RF) is calculated as follows:

1. For each decision tree in the forest, use the corresponding OOB samples to compute its OOB error, denoted $errOOB_1$;
2. Randomly add noise interference to feature/variable X in OOB sample (i.e., randomly change the sample's value at feature X), and again compute its OOB error, denoted as $errOOB_2$;
3. Let there be N_{tree} decision trees in the RF, then the VI of feature X can be expressed as:

$$\text{VI} = \frac{1}{N_{tree}}\sum\left(errOOB_2 - errOOB_1\right) \tag{2.24}$$

Measuring the variable importance in this way is considered from the point of view of the change in prediction accuracy: if the OOB accuracy of a feature changes substantially when a random noise perturbation is added to the feature, it means that the feature has a large impact on the prediction results, indicating a high feature importance.

Mean decreased Gini impurity

In tree-based ML models, the relative rank (i.e., depth) of a feature used as a decision node can be used to assess the relative importance of that feature with respect to the predictability of the target variable. This method relies on the Gini impurity criterion, which is used in the construction of decision trees of a random forest. Each time a feature is used to split a node, it reduces the impurity of that node. The sum of these reductions averaged across all trees in the forest, i.e.,

Mean decrease in impurity (MDI), indicates the feature's importance [82]. Features that result in large decreases in Gini impurity are considered more important because they help to create purer splits, leading to more accurate predictions.

In an RF, features that appear closer to the top of the decision trees influence the prediction for a larger portion of the input data. Therefore, the expected proportion of samples affected by these features can be used to estimate their relative importance. In implementation, the importance is further refined by combining the fraction of samples influenced by a feature with the reduction in impurity achieved by splitting on that feature, resulting in a normalized measure of its predictive power. The advantage of this method is computationally fast since the computation of importance can be done using only the intermediate values obtained during the random forest training (during the growth of each subtree in the forest). The implementation of RF feature importance in *scikit-learn* library [83] is MDI-based.

Permutation importance

Even though impurity-based feature importance (MDI) computed on tree-based models is one of the most commonly used methods, it suffers from two drawbacks that can lead to misleading conclusions. First, they are computed on statistics derived from the training dataset and hence do not necessarily provide insight into which features are most important to make good predictions on the held-out test set. Secondly, they tend to overemphasize features with high cardinality—those with many unique values.

Permutation feature importance [61] is a model-agnostic technique and can be used as an alternative to impurity-based feature importance that does not suffer from these flaws. Unlike impurity-based methods, which rely on the internal structure of tree-based models, permutation importance assesses the contribution of features by directly measuring their impact on model performance.

The key idea behind this method is to shuffle the values of a specific feature and observe the resulting change in the model's accuracy or another performance metric. If a feature is important, permuting its values will disrupt the relationship between the feature and the target variable, leading to a significant drop in model performance. Conversely, if a feature has little or no impact on predictions, the model's accuracy will remain largely unaffected by the permutation. By calculating the decrease in performance for each feature, permutation importance provides a ranking that reflects the contribution of each variable to the model's predictive power.

Although this method can overcome some biases related to feature cardinality or the specific mechanics of tree-based algorithms, it does have some limitations. Permutation importance requires multiple evaluations of the model by shuffling each feature and recalculating performance metrics. This can be computationally intensive, especially for large datasets or complex models with many features. It relies on the model's predictive accuracy, so if the model has poor perfor-

mance or is overfitted, permutation importance scores may be unreliable or misleading. Besides, when facing highly collinear features, permuting one feature might not significantly affect performance because the model can still get the same information from a correlated feature to make predictions, thus downplaying the actual contribution of the shuffled feature.

In summary, the assessment of variable importance is essential for understanding the contribution of each feature to a model's predictive performance. Each method has its strengths and limitations, and selecting the appropriate technique depends on the model and the specific goals of the analysis. For tree-based models like Random Forest, impurity-based measures may provide useful insights, while permutation importance offers a more general approach suitable for a wide range of models.

2.5.2 Sensitivity analysis

Sensitivity analysis is a crucial tool applied to multi-parametric mathematical models used to understand and assess the relative influence of each model parameter or model input, and their interrelation, on the output of the model [84]. It examines how the variation in the input parameters of a model impacts the outcome, thus helping to identify which inputs are the most influential. In essence, sensitivity analysis allows us to determine how sensitive the model's results are to changes in assumptions, data inputs, or parameters. By systematically varying input parameters, either individually or collectively, sensitivity analysis can reveal the extent to which uncertainty in the model's inputs contributes to uncertainty in its predictions. This process is particularly valuable in models with multiple uncertain parameters, as it can help prioritize parameters and inputs where data accuracy is most critical. Unfortunately, these analyses are very rarely applied in the context of ML.

The use of sensitivity analyses enhances the interpretability of models by ensuring that conclusions drawn from them are not overly reliant on uncertain or poorly understood inputs. This makes it a critical step in model validation, decision-making, and risk assessment, particularly in complex systems where model outcomes drive important real-world decisions.

Several methods exist for conducting sensitivity analysis, ranging from local methods, which involve changing one parameter at a time, to global methods, which explore the entire parameter space simultaneously. Local sensitivity analysis is often more straightforward but may overlook interactions between parameters, while global sensitivity analysis, though more computationally intensive, provides a more comprehensive understanding of parameter interactions and non-linear effects.

The screening methods are designed to give a coarse (not exploring the entire input space) and a computationally inexpensive assessment of the relative importance of input parameters [85, 86]. The Morris method is one of the most widely used screening techniques. We proposed an original

application of Morris screening methods in the context of the analysis of ML models, as presented in our published work [87] and following chapters.

Morris screening method

The Morris screening method [88] is a "one-step-at-a-time" (OAT) method in which only one input parameter is assigned with new values in each analysis. The key idea of the method is the estimation of Elementary Effects (EE) over multiple random sampling for each input parameter, allowing for the identification of influential parameters and possible interaction effects between parameters. Once the key parameters are identified through screening, if necessary, more refined global sensitivity analysis methods can be applied to these factors for deeper insights.

For a model $\mathcal{M}$ with K parameters, where $\boldsymbol{x} = [x_1, x_2, ..., x_K]$ is a vector of input with the K parameters and $y = \mathcal{M}(\boldsymbol{x})$ is the model output evaluated at point $\boldsymbol{x}$, the Morris method explores a sampled parameter space, a K-dimensional unit hypercube, with predefined supports for each parameter, subsequently divided into a uniform grid of points of p levels at which the model can be evaluated. On this evenly-spaced grid, the method then generates several random trajectories (R) allowing the parameter "jump" in the hypercube with a minimum step of $\frac{1}{p-1}$. The randomized trajectory design matrix is given in [88]. Each trajectory (r) evaluate the model $(K+1)$ times. From the starting point, which is a set of parameters randomly selected in the parameter space, each parameter will be varied or perturbed in turn and, particularly, in the form of a relay. That is to say that during this process, it does not return to the original starting point after perturbation but continues perturbing another dimension from the perturbed point, allowing efficient exploration of parameter space in given levels.

The estimation of Elementary Effects (EE) is associated with a given trajectory r in the parameter space, and it measures the changes in the output of the model $y = \mathcal{M}(\boldsymbol{x})$ when perturbing one parameter at a time. The EE of parameter k is defined as follows:

$$EE_k = \frac{\mathcal{M}(x_1, x_2, ..., x_k + \Delta, ..., x_K) - \mathcal{M}(x_1, x_2, ..., x_k, ..., x_K)}{\Delta} \tag{2.25}$$

where Δ is the "grid jump" represents a predefined variation such that $\boldsymbol{x} + \Delta$ is still in the specified domain of parameter space; Δ is a value in $[\frac{1}{p-1}, \frac{2}{p-1}, ..., 1 - \frac{1}{p-1}]$.

Consider that an n_R number of elementary effects (or replicates), associated with the k'th parameter have been sampled from the finite distribution of EE_k, which can be noted as $EE_k^{(r)}, r \in [1, R]$. Two essential statistical indicators of EE_k from the sampled trajectories can be calculated. The first is the absolute mean [85] that represents the linear effect of the variable x_k on the model output y, defined as

$$\mu_k^* = \frac{1}{n_R} \sum_{r=1}^{n_R} |EE_k^r| \tag{2.26}$$

The second is the standard deviation of the EE_k from all trajectories and it indicates the presence of non-linearity and/or interactions between parameter k and other parameters.

$$\sigma_k = \sqrt{\frac{1}{n_R}\sum_{r=1}^{n_R}\left(EE_k^r - \mu_k\right)^2} \tag{2.27}$$

A qualitative analysis can be performed by plotting the μ_k^* and σ_k of EE_k for each parameter on the $\mu^* - \sigma$ plane (as shown in Figure 2.15). As suggested by Morris [88], there are four possible categories of parameter importance:

- Parameters with **non-influential** overall effects on the model output: relatively low values of both μ_k^* and σ_k. The ones are often clustered closer to the origin with a pronounced boundary.
- Parameters with **linear and/or additive but non-interacting** effects: relatively large value of μ_k^* and relatively small value of σ_k.
- Parameters with **non-linear and/or interacting** effects: relatively small value of μ_k^* and relatively large value of σ_k.
- Parameters with **less influential** effects: moderate values for both μ_k^* and σ_k.

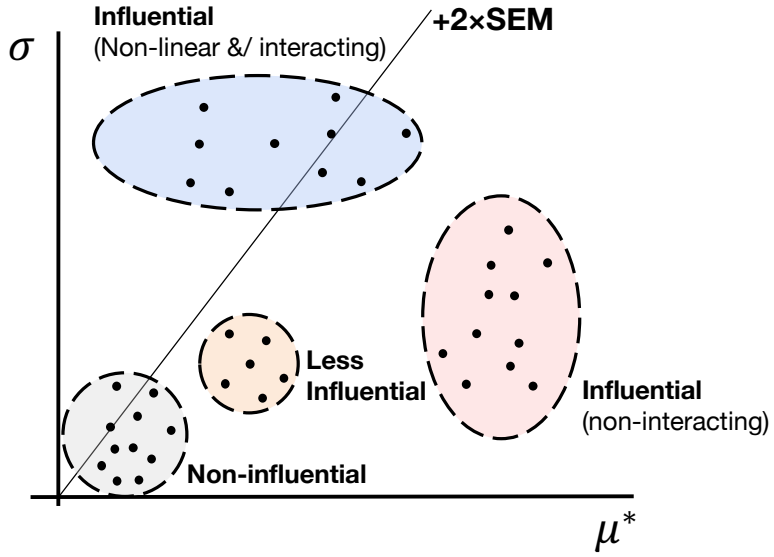

Figure 2.15: Illustration of qualitative analysis on parameter influence based on Elementary Effects of the Morris method. Each input parameter (feature) is represented as a point on the $\mu^* - \sigma$ plane, whose relative position offers information on the type of effect the parameter has on the output of the model in question. *SEM*: standard error of the mean.

By examining the μ^* of EE_k, we can estimate relative importance ranking among parameters and, in turn, make a deeper exploration of the most influential parameters on the model prediction outcomes or screen out non-influential ones for model compactness. Meanwhile, the analysis

of σ gives an indication of the interactions between the inputs, which offers an identification of simultaneous variations between parameters.

2.6 Conclusion

In conclusion, this chapter has outlined the key methodologies and tools employed throughout the dissertation, starting with an overview of the CARESS-Premi project and its data acquisition process. The proposed data processing pipeline has been detailed, providing the foundation for subsequent analyses. Additionally, we have introduced the statistical techniques, machine learning algorithms, and interpretability analyses that underpin the research studies presented in the following chapters, establishing a comprehensive framework for the investigation of neonatal care in the NICU setting.

BIBLIOGRAPHY

[1] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," IEEE Transactions on Biomedical Engineering, vol. BME-32, no. 3, pp. 230–236, 1985. [Online]. Available: http://ieeexplore.ieee.org/document/4122029/

[2] S. Osowski and T. H. Linh, "Ecg beat recognition using fuzzy hybrid neural network," IEEE Transactions on Biomedical Engineering, vol. 48, no. 11, pp. 1265–1271, 2001.

[3] Z. Zidelmal, A. Amirou, D. Ould-Abdeslam, A. Moukadem, and A. Dieterlen, "QRS detection using S-Transform and Shannon energy," Computer methods and programs in biomedicine, vol. 116, no. 1, pp. 1–9, 2014.

[4] C. Li, C. Zheng, and C. Tai, "Detection of ECG characteristic points using wavelet transforms," IEEE Transactions on biomedical Engineering, vol. 42, no. 1, pp. 21–28, 1995.

[5] D. Benitez, P. Gaydecki, A. Zaidi, and A. Fitzpatrick, "The use of the Hilbert transform in ECG signal analysis," Computers in biology and medicine, vol. 31, no. 5, pp. 399–406, 2001.

[6] Z. Zidelmal, A. Amirou, M. Adnane, and A. Belouchrani, "QRS detection based on wavelet coefficients," Computer methods and programs in biomedicine, vol. 107, no. 3, pp. 490–496, 2012.

[7] R. Rani, V. Chouhan, and H. Sinha, "Automated detection of QRS complex in ECG signal using wavelet transform," International Journal of Computer Science and Network Security (IJCSNS), vol. 15, no. 1, p. 1, 2015.

[8] M. U. Zahid, S. Kiranyaz, T. Ince, O. C. Devecioglu, M. E. H. Chowdhury, A. Khandakar, A. Tahir, and M. Gabbouj, "Robust R-peak detection in low-quality Holter ECGs using 1D convolutional neural network," IEEE Transactions on Biomedical Engineering, vol. 69, no. 1, pp. 119–128, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9451595/

[9] M. Gabbouj, S. Kiranyaz, J. Malik, M. U. Zahid, T. Ince, M. E. Chowdhury, A. Khandakar, and A. Tahir, "Robust peak detection for Holter ECGs by self-organized operational neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 9363–9374, 2022.

[10] M. Doyen, D. Ge, A. Beuchée, G. Carrault, and A. I. Hernández, "Robust, real-time generic detector based on a multi-feature probabilistic method," PLoS ONE, vol. 14, no. 10, p. e0223785, 2019.

[11] A. Beuchee, A. I. Hernandez, C. Duvareille, D. Daniel, N. Samson, P. Pladys, and J.-P. Praud, "Influence of hypoxia and hypercapnia on sleep state-dependent heart rate variability behavior in newborn lambs," Sleep, vol. 35, no. 11, pp. 1541–1549, 2012.

[12] R. McCraty and F. Shaffer, "Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk," Global advances in health and medicine, vol. 4, no. 1, pp. 46–61, 2015.

[13] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," Frontiers in public health, vol. 5, p. 258, 2017.

[14] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger et al., "Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology," Circulation, vol. 93, no. 5, pp. 1043–1065, 1996.

[15] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability," Frontiers in psychology, vol. 5, p. 1040, 2014.

[16] D. R. Kommers, R. Joshi, C. van Pul, L. Atallah, L. Feijs, G. Oei, S. B. Oetomo, and P. Andriessen, "Features of heart rate variability capture regulatory changes during kangaroo care in preterm infants," The journal of pediatrics, vol. 182, pp. 92–98, 2017.

[17] B. P. Kovatchev, L. S. Farhy, H. Cao, M. P. Griffin, D. E. Lake, and J. R. Moorman, "Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome," Pediatric research, vol. 54, no. 6, pp. 892–898, 2003.

[18] A. Bauer, J. W. Kantelhardt, A. Bunde, P. Barthel, R. Schneider, M. Malik, and G. Schmidt, "Phase-rectified signal averaging detects quasi-periodicities in non-stationary data," Physica A: Statistical Mechanics and its Applications, vol. 364, pp. 423–434, 2006.

[19] Electrophysiology, Task Force of the European Society of Cardiology the North American Society of Pacing, "Heart rate variability: standards of measurement, physiological interpretation, and clinical use," Circulation, vol. 93, no. 5, pp. 1043–1065, 1996.

[20] U. Chatow, S. Davidson, B. L. Reichman, and S. Akselrod, "Development and maturation of the autonomic nervous system in premature and full-term infants using spectral analysis of heart rate fluctuations," Pediatric research, vol. 37, no. 3, pp. 294–302, 1995.

[21] S. Cardoso, M. J. Silva, and H. Guimarães, "Autonomic nervous system in newborns: a review based on heart rate variability," Child's Nervous System, vol. 33, pp. 1053–1063, 2017.

[22] S. L. Marple Jr, Digital spectral analysis. Courier Dover Publications, 2019.

[23] S. M. Kay and S. L. Marple, "Spectrum analysis—a modern perspective," Proceedings of the IEEE, vol. 69, no. 11, pp. 1380–1419, 1981.

[24] T. Cokelaer and J. Hasch, "'spectrum': spectral analysis in python," Journal of Open Source Software, vol. 2, no. 18, p. 348, 2017.

[25] T. Kuusela, "Methodological aspects of heart rate variability analysis," Heart rate variability (HRV) signal analysis: Clinical applications, pp. 10–42, 2013.

[26] J. Piskorski and P. Guzik, "Filtering poincare plots," Computational methods in science and technology, vol. 11, no. 1, pp. 39–48, 2005.

[27] A. B. Ciccone, J. A. Siedlik, J. M. Wecht, J. A. Deckert, N. D. Nguyen, and J. P. Weir, "Reminder: RMSSD and SD1 are identical heart rate variability metrics," Muscle & nerve, vol. 56, no. 4, pp. 674–678, 2017.

[28] N. Lippman, K. M. Stein, and B. B. Lerman, "Comparison of methods for removal of ectopy in measurement of heart rate variability," American Journal of Physiology-Heart and Circulatory Physiology, vol. 267, no. 1, pp. H411–H418, 1994.

[29] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," American journal of physiology-heart and circulatory physiology, vol. 278, no. 6, pp. H2039–H2049, 2000.

[30] D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman, "Sample entropy analysis of neonatal heart rate variability," American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, vol. 283, no. 3, pp. R789–R797, 2002.

[31] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," Chaos: an interdisciplinary journal of nonlinear science, vol. 5, no. 1, pp. 82–87, 1995.

[32] N. Iyengar, C. Peng, R. Morin, A. L. Goldberger, and L. A. Lipsitz, "Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics," American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, vol. 271, no. 4, pp. R1078–R1084, 1996.

[33] T. Nakamura, H. Horio, S. Miyashita, Y. Chiba, and S. Sato, "Identification of development and autonomic nerve activity from heart rate variability in preterm infants," Biosystems, vol. 79, no. 1-3, pp. 117–124, 2005.

[34] M. Kobayashi and T. Musha, "1/f fluctuation of heartbeat period," IEEE Transactions on Biomedical Engineering, no. 6, pp. 456–457, 1982.

[35] D. Giavarina, "Understanding Bland Altman analysis," Biochemia Medica, vol. 25, no. 2, pp. 141–151, 2015.

[36] K. Pearson, "Notes on the history of correlation," Biometrika, vol. 13, no. 1, pp. 25–45, 1920.

[37] C. Spearman, "The proof and measurement of association between two things." 1961.

[38] C. F. Dietrich, Uncertainty, calibration and probability: the statistics of scientific and industrial measurement. Routledge, 2017.

[39] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," Nature Methods, vol. 17, pp. 261–272, 2020.

[40] R. L. Plackett, "Studies in the History of Probability and Statistics. XXIX: The discovery of the method of least squares," Biometrika, vol. 59, no. 2, pp. 239–251, 1972.

[41] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings. Springer, 2000, pp. 298–372.

[42] K. W. Vugrin, L. P. Swiler, R. M. Roberts, N. J. Stucky-Mack, and S. P. Sullivan, "Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies," Water Resources Research, vol. 43, no. 3, 2007.

[43] D. G. Altman and J. M. Bland, "Measurement in medicine: The analysis of method comparison studies," Journal of the Royal Statistical Society Series D: The Statistician, vol. 32, no. 3, pp. 307–317, 1983.

[44] J. Martin Bland and DouglasG. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," The Lancet, vol. 327, no. 8476, pp. 307–310, 1986.

[45] K. M. Ho, "Using linear regression to assess dose-dependent bias on a Bland-Altman plot," Journal of Emergency and Critical Care Medicine, vol. 2, no. 8, 2018.

[46] H. Abdi and L. J. Williams, "Principal component analysis," Wiley interdisciplinary reviews: computational statistics, vol. 2, no. 4, pp. 433–459, 2010.

[47] M. Geun Kim, "Multivariate outliers and decompositions of mahalanobis distance," Communications in statistics-theory and methods, vol. 29, no. 7, pp. 1511–1526, 2000.

[48] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," Data & Knowledge Engineering, vol. 63, no. 2, pp. 503–527, 2007.

[49] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 444–452.

[50] J. M. Helm, A. M. Swiergosz, H. S. Haeberle, J. M. Karnuta, J. L. Schaffer, V. E. Krebs, A. I. Spitzer, and P. N. Ramkumar, "Machine learning and artificial intelligence: definitions, applications, and future directions," Current reviews in musculoskeletal medicine, vol. 13, pp. 69–76, 2020.

[51] A. L. Samuel, "Some studies in machine learning using the game of checkers," IBM Journal of research and development, vol. 3, no. 3, pp. 210–229, 1959.

[52] A. Burkov, The hundred-page machine learning book. Andriy Burkov Quebec City, QC, Canada, 2019, vol. 1.

[53] S. J. Russell and P. Norvig, Artificial intelligence: a modern approach. Pearson, 2016.

[54] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255–260, 2015.

[55] H. Zhang, "The optimality of naive Bayes," American Association for Artificial Intelligence, vol. 1, no. 2, p. 3, 2004.

[56] E. Frank, L. Trigg, G. Holmes, and I. H. Witten, "Naive Bayes for regression," Machine Learning, vol. 41, pp. 5–25, 2000.

[57] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," Journal of biomedical informatics, vol. 35, no. 5-6, pp. 352–359, 2002.

[58] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, pp. 81–106, 1986.

[59] ——, C4. 5: programs for machine learning. Elsevier, 2014.

[60] L. Breiman, Classification and regression trees. Routledge, 2017.

[61] ——, "Random forests," Machine learning, vol. 45, pp. 5–32, 2001.

[62] ——, "Bagging predictors," Machine learning, vol. 24, pp. 123–140, 1996.

[63] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[64] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, pp. 1189–1232, 2001.

[65] S. Haykin, Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.

[66] A. F. Agarap, "An architecture combining convolutional neural network (CNN) and support vector machine (SVM) for image classification," arXiv preprint arXiv:1712.03541, 2017.

[67] T. K. Lee, W. J. Baddar, S. T. Kim, and Y. M. Ro, "Convolution with logarithmic filter groups for efficient shallow cnn," in MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24. Springer, 2018, pp. 117–129.

[68] F. Lei, X. Liu, Q. Dai, and B. W.-K. Ling, "Shallow convolutional neural network for image classification," SN Applied Sciences, vol. 2, no. 1, p. 97, 2020.

[69] H. Wu and J.-T. Zhang, Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches. John Wiley & Sons, 2006.

[70] H. J. Seltman, "Experimental design and analysis," https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/, p. 357–378, 2012, Accessed: 2024-09-18.

[71] UCLA: Statistical Consulting Group, "Introduction to linear mixed models," https://stats.oarc.ucla.edu/sas/modules/introduction-to-the-features-of-sas/, 2021, Accessed: 2024-09-18.

[72] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed effects regression trees for clustered data," Statistics & probability letters, vol. 81, no. 4, pp. 451–459, 2011.

[73] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the royal statistical society: series B (methodological), vol. 39, no. 1, pp. 1–22, 1977.

[74] G. J. McLachlan and T. Krishnan, The EM algorithm and extensions. John Wiley & Sons, 2007.

[75] R. J. Sela and J. S. Simonoff, "RE-EM trees: a data mining approach for longitudinal and clustered data," Machine learning, vol. 86, pp. 169–207, 2012.

[76] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed-effects random forest for clustered data," Journal of Statistical Computation and Simulation, vol. 84, no. 6, pp. 1313–1328, 2014.

[77] K. Jamieson and A. Talwalkar, "Non-stochastic best arm identification and hyperparameter optimization," in Artificial intelligence and statistics. PMLR, 2016, pp. 240–248.

[78] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in International conference on machine learning. PMLR, 2013, pp. 115–123.

[79] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-Sklearn: Automatic hyperparameter configuration for scikit-learn." in Scipy, 2014, pp. 32–37.

[80] T. Bäck and H.-P. Schwefel, "An overview of evolutionary algorithms for parameter optimization," Evolutionary computation, vol. 1, no. 1, pp. 1–23, 1993.

[81] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer, 2009, vol. 2.

[82] G. Louppe, "Understanding random forests: From theory to practice," arXiv preprint arXiv:1407.7502, 2014.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[84] A. Saltelli, S. Tarantola, and F. Campolongo, "Sensitivity analysis as an ingredient of modeling," Statistical science, pp. 377–395, 2000.

[85] F. Campolongo, J. Cariboni, and A. Saltelli, "An effective screening design for sensitivity analysis of large models," Environmental modelling & software, vol. 22, no. 10, pp. 1509–1518, 2007.

[86] F. Campolongo, A. Saltelli, and J. Cariboni, "From screening to quantitative sensitivity analysis. a unified approach," Computer physics communications, vol. 182, no. 4, pp. 978–988, 2011.

[87] M. Chen and A. Hernández, "Towards an explainable model for sepsis detection based on sensitivity analysis," IRBM, vol. 43, no. 1, pp. 75–86, 2022.

[88] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," Technometrics, vol. 33, no. 2, pp. 161–174, 1991.

CHAPTER

3 Model-based Characterization of Bilirubin Dynamics in Preterm Infants

Hyperbilirubinemia in preterm infants, as introduced in Section 1.3.1, remains a significant clinical challenge in NICU settings. This condition, if left untreated, can lead to severe neurodevelopmental disorders such as kernicterus [1–4]. The monitoring and management of neonatal hyperbilirubinemia is particularly complex in preterm infants due to their physiological immaturity and the highly variable bilirubin dynamics.

In the following two chapters, we investigate two aspects of the management of neonatal hyperbilirubinemia (NHB), concerning model-based characterization of total serum bilirubin dynamics and knowledge-based non-invasive estimation of TSB using mixed-effects machine learning models. Together, these studies aim to offer a comprehensive approach to tackling the clinical challenges of hyperbilirubinemia in preterm infants and to pave the way for more effective and patient-specific clinical interventions.

This chapter, strongly based on our publication [5], presents the first aspect, on the validation of a patient-specific exponential decay model to characterize the natural and long-term dynamics of total serum bilirubin (TSB) concentrations in preterm infants born between 24 to 32 weeks of gestation, and to study the model parameters as potential biomarkers for detecting associated morbidities.

We first present a literature review of studies describing the evolution of bilirubin levels during the early postnatal period in term and preterm infants. Then we propose the study data and methodology concerning a deterministic mathematical approach that describes bilirubin kinetics over the initial weeks of life, which is an unprecedented long-term view of TSB evolution in this vulnerable population. In addition, the potential of the proposed model and its parameters as new indicators of high-risk clinical events in the NICU settings are also explored.

3.1 Introduction

Total serum bilirubin (TSB) or Transcutaneous bilirubin (TcB) levels reflect the balance between bilirubin production and its elimination. It is of great importance to understand the natural course of the TSB evolution during the first weeks of life in preterm infants since this tiniest population is at higher risk of hyperbilirubinemia and its complications such as Bilirubin-induced neurologic dysfunction (BIND), Necrotizing enterocolitis (NEC), infections, etc. A comprehensive understanding of bilirubin dynamics in the neonatal period is of significant value in distinguishing between normal and abnormal conditions. Furthermore, it plays a pivotal role in optimizing care and eventually enhancing the outcomes of newborns during their critical early weeks.

In the literature, rich studies have been dedicated to describing the TSB or TcB evolution in the early postnatal period, i.e., within the first 72 or 96 hours of life [6–8].

An early study by Bhutani et al. [6] developed a predictive percentile-based bilirubin nomogram from hour-specific pre-discharge and post-discharge TSB values of 2,840 healthy term and near-term newborns, with mean Gestational age (GA) of 38.7$\pm$1.3 weeks and birth weight of 3,318$\pm$457 grams, measured between 18 and 72 hours of postnatal age. The nomogram was assessed for its ability to predict the risk of subsequent clinically significant hyperbilirubinemia as high risk ($\geq$95th percentile), intermediate-risk (40th–95th percentiles), and low-risk (<40th percentiles) (Figure 3.1). This work was the foundation for a wide range of national guidelines for managing NHB including the widely practiced AAP recommendations [2, 9].

DeLuca et al. [7] conducted a systematic review to study a mathematical model of bilirubin kinetics describing the natural course of physiologic jaundice and the rate of rise (ROR) of TcB levels. Among the included subjects, i.e., healthy neonates with at least 35 weeks of gestational age, the ROR of TcB described by a quadratic equation showed a general decrease with increasing postnatal age until reaching a plateau at about 96 hours of life. Their results suggested that a higher bilirubin ROR indicates a relatively high risk for subsequent hyperbilirubinemia in neonates and should lead to close monitoring and additional care.

Similarly while more globally, Kaplan et al. [8] carried out a more extensive review on a global sample of 20 TcB nomograms from 19 worldwide studies targeting newborns $\geq$35 weeks of gestation. From the reviewed nomograms, the authors approximated the natural history of hyperbilirubinemia during the first 5 days of life in apparently normal and predominantly breastfed neonates and thereby depicted a TcB nomogram (as shown in Figure 3.2) that reflects the natural history of early NHB. It shows an increase in TcB during the first 3 postnatal days, and the TcB levels peaked or plateaued between the 3rd and 4th days. The TcB trajectories in the proposed nomogram validated the evolution patterns of TSB and TcB in previous studies [10, 11].

In addition to the studies that have mainly focused on full-term and late-preterm infants, recent works [12–15], as a complement, have filled the gap of less attention paid to the evolutionary

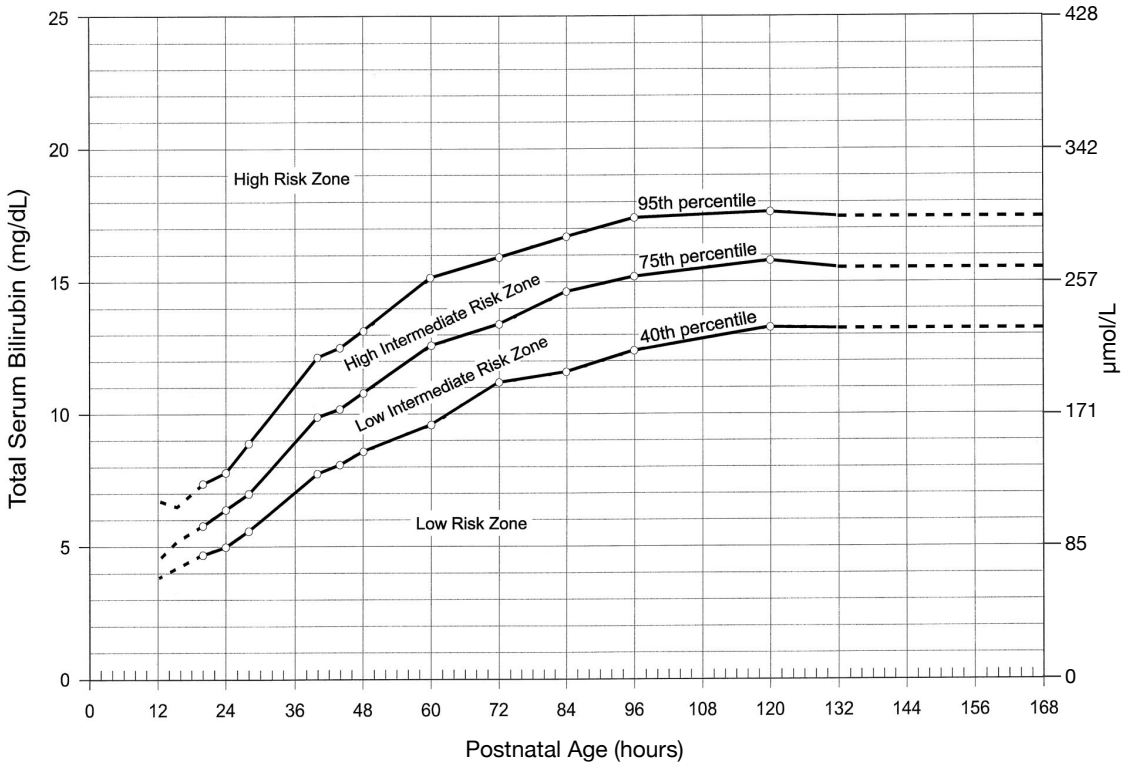

Figure 3.1: Nomogram for designation of risk in 2,840 well newborns at 36 or more gestational weeks with a birth weight of 2,000 grams or more **or** 35 or more gestational weeks and birth weight of 2,500 grams or more based on the hour-specific serum bilirubin values. The high-risk zone is designated by the 95th percentile track. The intermediate-risk zone is subdivided into upper- and lower-risk zones by the 75th percentile track. The low-risk zone has been electively and statistically defined by the 40th percentile track. Dotted extensions are based on <300 TSB values/epoch. This nomogram should not be used to represent the natural history of NHB. *(Adapted from [6, 9]. Reproduced with permission from Journal Pediatrics, Vol. 103, Page(s) 6-14, Copyright ©1999 by the AAP.)*

pattern of bilirubin concentrations in preterm infants <35 gestational weeks and/or those with low birth weight, who are at high risk for subsequent hyperbilirubinemia.

Maisels et al. [12] introduced recommendations to manage and treat hyperbilirubinemia in preterm infants born at <35 weeks' gestation to be complementary of the guidelines for the management of hyperbilirubinemia in the neonates ≥35 weeks gestation. However, it is largely based on expert opinion rather than solid evidence potentially resulting in increased use of phototherapy in this population [16].

Hahn et al. [13] examined the natural course of hour-specific TSB levels during the first 72 hours of life before the initiation of phototherapy in 483 Very low birth weight (VLBW) preterm newborns and modeled the dynamics against postnatal age as a square root curve (r=0.843, p

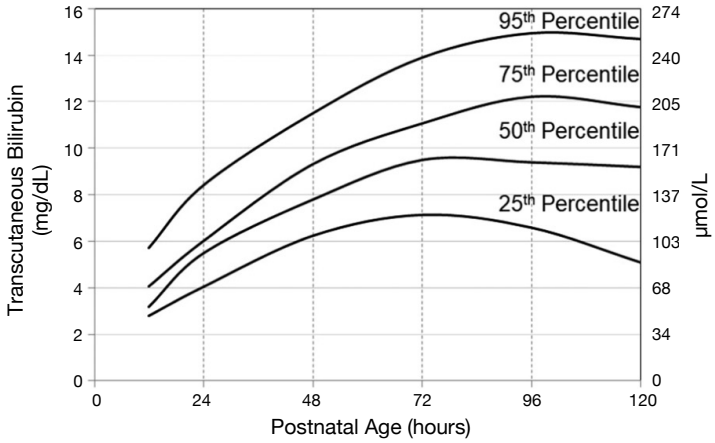

Figure 3.2: Nomogram constructed from pooled transcutaneous bilirubin readings. The nomogram was derived from 20 nomograms from 12 different countries, depicting the approximate natural history of post-natal bilirubinemia in normal, predominantly breastfed newborns ≥35 weeks of gestation.
(Adapted from [8]. Copyright ©2021. Reproduced with permission from Springer Nature.)

<0.001). They also evaluated the ROR of TSB and confirmed that it could be an indicator for both initiating timing and duration of phototherapy, that is, the subgroup displayed rapidly rising TSB levels ≥95th percentile (>0.25 mg/dL/h) had received significant earlier and longer phototherapy than a subgroup of "slow risers" (mean ROR <0.25 mg/dL/h). Besides, TSB appeared to rise more rapidly in infants of low gestational age, low birth weight and low 5-minute Apgar scores.

Jegathesan et al. [14] generated hour-specific pre-treatment TSB percentile-based nomograms among 6,143 pre-treatment TSB measures from 2,549 infants which was expected to inform how bilirubin is described in preterm newborns born at $29^{0/7}$-$35^{6/7}$ gestational weeks. In the main nomogram contributed by overall infants in the study, similar patterns of ROR were observed: the ROR of TSB declined with advancing hours after birth (Figure 3.3a). Further nomograms were developed by gestational age groups ($29^{0/7}$-$32^{6/7}$ and $33^{0/7}$-$35^{6/7}$ weeks' gestation) and by subsequent receipt of phototherapy. One of their results indicated that infants born at 29-32 weeks' gestation had a significantly lower but earlier peak in mean pre-treatment TSB levels, compared to the relatively mature infant group born at 33-35 weeks of gestation. As for the phototherapy recipients, within the first 72 hours, they had a higher mean peak in TSB concentrations (145.7 µmol/L versus 132.1 µmol/L; p <0.01) that occurred significantly earlier (38.3 hours versus 50.8 hours; p <0.01).

Following the above study, Jegathesan and his team moved a step further to generate percentile-based pre-phototherapy TSB levels in a cohort of 642 extremely preterm infants born at $24^{0/7}$-$28^{6/7}$ weeks' gestation [15] (Figure 3.3b). In addition, they compared 24-hour pre-phototherapy TSB percentiles with existing consensus-based phototherapy guidelines [12] and pointed out the high

frequency of phototherapy use and provided a contemporary understanding of pre-phototherapy TSB levels in extremely preterm infants.

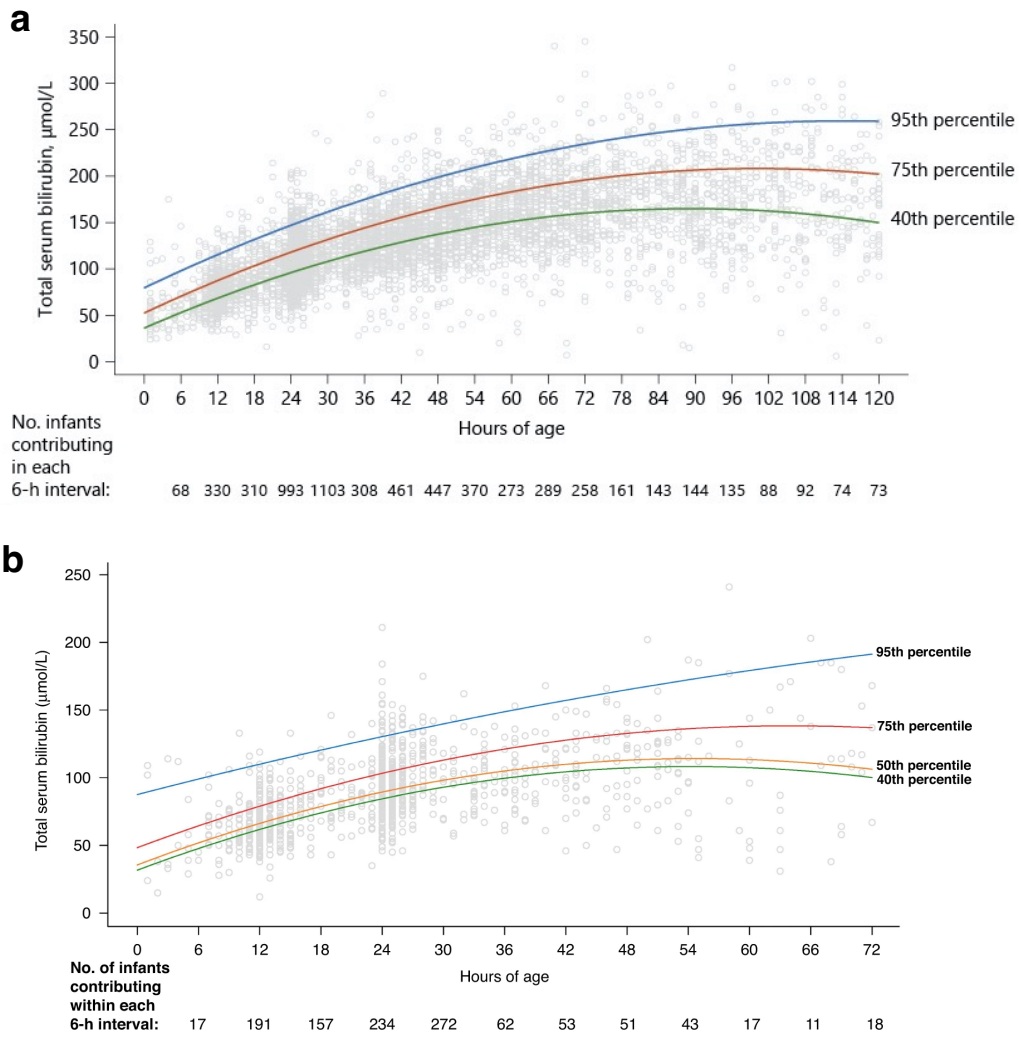

Figure 3.3: Hour-specific pre-treatment total serum bilirubin percentile-based curves among (**a**) preterm infants born at $29^{0/7}$-$35^{6/7}$ weeks' gestation (n=2,549), and (**b**) extremely preterm infants born at $24^{0/7}$-$28^{6/7}$ weeks' gestation (n=642). Pre-treatment TSB levels refer to TSB levels prior to phototherapy among those administered phototherapy and any TSB levels among those not administered phototherapy.
((a) Adapted from [14]. Copyright ©2021 Karger Publishers, Basel, Switzerland.
(b) Adapted from [15]. Reproduced with permission from Springer Nature. Copyright ©2022 Springer Nature.)

However, despite increased research efforts in the cohorts of preterm infants, there remains

a paucity of evidence on the long-course natural evolution of bilirubin, especially after treatment, in preterm and extremely preterm infants, who are more likely to face potentially life-threatening conditions such as rebound hyperbilirubinemia, infection and adverse neurological dysfunctions, etc. Continuous and prolonged TSB monitoring is recommended for all preterm infants, as there have been reports of a late-onset increase in TSB up to 110 hours old and related delayed hyperbilirubinemia [14, 17]. Thus, studying how natural bilirubin levels in premature infants evolve and develop as they age over an extended neonatal period (e.g., weeks after birth), will provide important and comprehensive new perspectives into understanding the long-term bilirubin dynamics.

On the other hand, in the context of precision medicine, the focus on individualized diagnosis and medical interventions has led to growing interest among healthcare professionals and researchers in patient-specific modeling. Patient-specific modeling facilitates personalized characterization of bilirubin dynamics in preterm infants, who exhibit significant variability regarding the level of immaturity and many other relevant clinical factors. By accounting for these individual differences, patient-specific models can optimize the diagnosis and interventions. Furthermore, the integration of patient-specific modeling into clinical practice may foster understanding and managing the complexities of neonatal hyperbilirubinemia, thereby providing more targeted care and support.

Physiologically and naturally, the ability of preterm infants to metabolize and excrete bilirubin matures as aging after birth. After the period of physiological neonatal jaundice (details refer to Section 1.3.1), as the liver matures, its ability to metabolize bilirubin improves, leading to a progressive reduction in the accumulation of bilirubin in the blood, i.e., TSB.

An important aspect of the metabolic pathway of bilirubin excretion is the bilirubin conjugating capacity of the liver, mediated by hepatic bilirubin UDP-glucuronosyltransferase activity [18, 19]. It was reported that during the perinatal period (Figure 3.4a), this enzyme activity is observed to gradually but significantly increase, in fetuses, premature and full-term infants born between 30 and 40 weeks of gestation and survived less than 7 days of life, from 0.1% to 1.0% of the values found in the adult liver. After birth, the activity begins to increase at an exponential rate until it reaches the adult value by 14 weeks of age, after which it remains constant until adulthood. Figure 3.4b further demonstrated that the transferase activity increases after birth in premature infants who survived for 8 days to 28 days of life.

Based on the above, in this study, we thus focus on a population of preterm infants and hypothesize that the natural dynamics of TSB concentrations from three postnatal days could be quantified and modeled through an exponential decay. Furthermore, the parameters of such a model, once fitted to reflect the TSB dynamics at the individual level, might be used to characterize the maturation and functionality of the bilirubin conjugation pathway.

The primary objective of this study is to investigate the characteristics of the age-related dy-

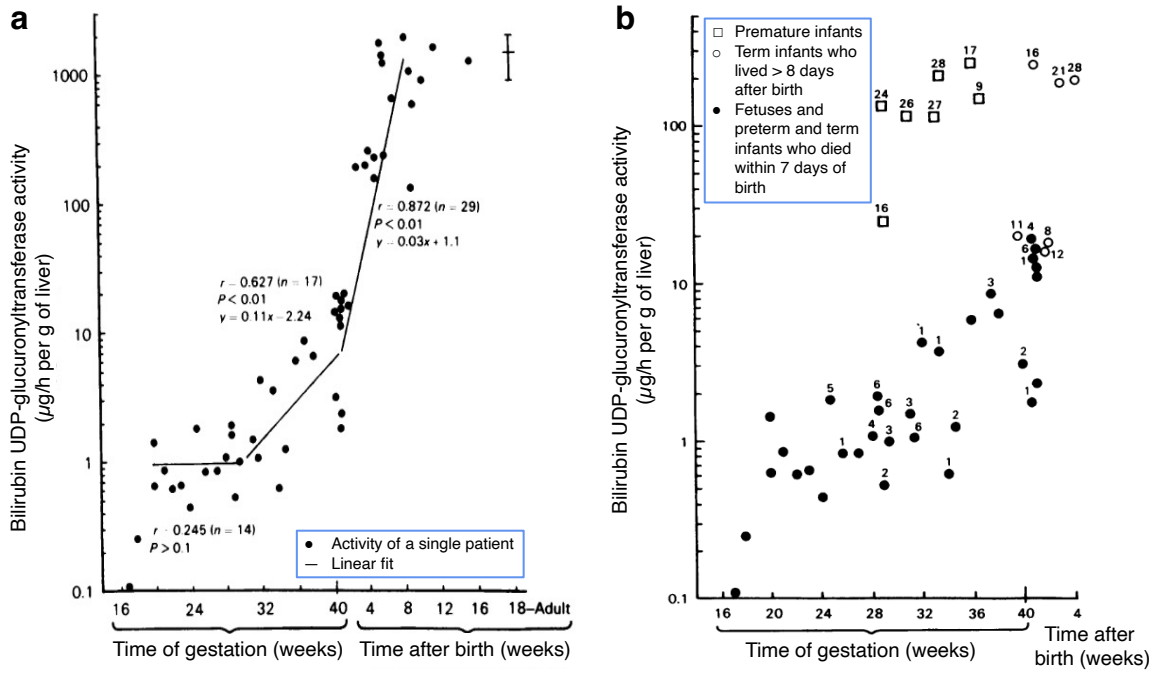

Figure 3.4: (**a**) Developmental pattern in human hepatic bilirubin UDP-glucuronosyltransferase activities. The enzyme activities were plotted against weeks of gestation and weeks after birth on a semi-logarithmic scale. (**b**) Effect of premature birth on the development of human hepatic bilirubin UDP-glucuronosyltransferase activities. The enzyme activities were plotted against weeks of gestation and weeks after birth on a semi-logarithmic scale. Preterm delivery, irrespective of gestational age, evokes an early increase in transferase activities, equal in rate to the normal postnatal increase. The number shown beside the symbols represents the age (days) at which death occurred.
(Adapted from [18].)

namics of TSB levels during the early postnatal period, aiming to track and anticipate developmental patterns in this population. Our secondary objective is to validate the parameters of models as potential biomarkers for detecting relevant comorbidities when bilirubin evolutionary trends diverge from expected decays. Studying how natural bilirubin levels in premature infants evolve and develop over the first few days to weeks after birth as they age will provide important and comprehensive new perspectives into understanding the long-term bilirubin dynamics.

3.2 Materials and Methods

3.2.1 Population and inclusion criteria

Within the database carried out by the CARESS-Premi clinical study (details refer to Section 2.1.3), the dominant center—*University Hospital Center of Rennes* (*CHU Rennes*)—consisting of 414 patients is considered in this study. Among those, 373 infants who have undergone bilirubin-related conditions during the follow-ups are therefore eligible for analysis.

In the first stage, infants with fewer than 4 measurements are excluded. This exclusion is based on the consideration that a reasonable number of samples is required to fit a robust model. Then, we examine three conditions within the included population and consider them for exclusion. The first category comprises patients with infrequent monitoring and those with adjacent TSB measurements taken more than 10 days apart. This exclusion was necessary since sparse sampling might fail to capture possible intermediate abnormal fluctuations in TSB levels, leading to inaccurate or overly optimistic representation of the bilirubin decay in the modeling fitting process. The other two categories are Exchange transfusion (ET) recipients under NICE guidelines, and Phototherapy (PT) recipients as documented in the database. Patients exhibiting infrequent TSB monitoring were directly excluded from the population. For ET and PT recipients, patients with at least 4 remaining measurements after removal of samples taken during treatment were retained for further analysis. These exclusions ensure that the models effectively characterize the natural evolution of TSB levels, unaffected by treatments that may alter TSB dynamics.

No blood sample was specifically collected for the study; instead, blood samples used to determine total serum bilirubin levels were collected for routine clinical care and adhered to standard care criteria.

In addition to the Postnatal age (PNA) and TSB measurements, relevant clinical characteristics including demographic, maternal, laboratory tests such as C-reactive protein (CRP), treatments, and short-term outcome data were extracted from the CARESS-Premi database. Necrotizing enterocolitis (NEC) was defined as a grade II-a or higher according to the modified Bell's staging criteria [20].

The Z-scored birth weights (BW) were additionally calculated according to the sex-specific Fenton 2013 preterm growth charts developed by meta-analyses of about 4 million births from six countries [21]. It represents the difference between the observed BW and the expected BW for the subject's gestational age and sex, considering the variability in growth. A Z-scored BW of zero indicates that the neonate's BW is equivalent to the mean weights of the reference population at the same GA and sex, suggesting an appropriate growth for her/his GA. A Z-scored BW below -2 or -3 typically signifies a neonate who is classified as "Small for Gestational Age", while a z-score above +2 or +3 indicates a neonate as "Large for Gestational Age".

3.2.2 General TSB decay models

An overall physiological downward trend in Total serum bilirubin (TSB) levels after birth, following physiological neonatal jaundice and some fluctuations due to various events, is observed among premature newborns [18, 19, 22]. To characterize the decay trend, we first compared two general models: a linear regression model and a basic exponential decay model. These mathematical models were fitted to all available TSB measurements from all relevant infants, sorted by postnatal age (PNA), without accounting for individual variability.

A simple linear regression model assumes that the TSB levels decrease linearly over time, expressed as:

$$\boldsymbol{TSB}(t) = A + B\boldsymbol{t} + \epsilon \tag{3.1}$$

where $\boldsymbol{t}$ is the PNA in days when the corresponding TSB level (µmol/L) was measured, A is the intercept, B is the slope, and ϵ is the error term capturing deviations of the fitted line from raw samples.

A basic exponential decay model posits that the TSB levels exponentially decline over time, with the rate of decay proportional to the current value of TSB, formulated as:

$$\boldsymbol{TSB}(t) = A \times exp(-B\boldsymbol{t}) + C + \epsilon \tag{3.2}$$

where $\boldsymbol{t}$ is the PNA in days when the corresponding TSB level (µmol/L) was measured, A represents the initial TSB level, B is the decay factor indicating the rate of decline, C is the constant term representing the baseline level that TSB asymptotically approaches, and ϵ is the error term. Note that the negative sign before the parameter B guarantees the decay rather than the increase of the function.

The linear regression is conducted using the Ordinary least squares (OLS) method implemented in *statsmodels* library [23] in Python; and the general exponential decay model is fitted using the non-linear OLS method implemented in *SciPy* [24] library in Python.

3.2.3 Patient-specific TSB exponential decay model

Recognizing the need to accommodate inter-individual variability in TSB dynamics among premature infants, we further proposed a mathematically deterministic function to model patient-specific exponential decay patterns. The model, namely, $g(\cdot)$, accounts for individual differences by fitting the evolution of TSB levels to each infant individually, adjusting for the infant's GA and

PNA. For each subject i, the patient-specific TSB exponential decay model is given by:

$$\begin{aligned} \boldsymbol{TSB}_i(t) &= g(\boldsymbol{p}_i(t); \boldsymbol{S}_i) + \boldsymbol{\epsilon}_i \\ &= g(\boldsymbol{p}_i(t); A_i, B_i, C_i, GA_i, Tc_i) + \boldsymbol{\epsilon}_i \\ &= A_i \times exp\left(-B_i\left[\boldsymbol{p}_i(t) + GA_i + Tc_i\right]\right) + C_i + \boldsymbol{\epsilon}_i \end{aligned} \tag{3.3}$$

where $\boldsymbol{p}_i(t)$ refers to a sequence of PNA (days) as the independent variable, $\boldsymbol{TSB}_i(t)$ is the TSB concentrations (µmol/L) observed at $\boldsymbol{p}_i$ corresponding time t as the dependent variable, and $\boldsymbol{\epsilon}_i$ the associated modeling error. The parameters set $\boldsymbol{S}_i$ is specific to each infant (i), including five parameters: A_i, B_i, C_i, GA_i and Tc_i. Parameters A_i, B_i and C_i determine the fitting curve's morphology and control how it evolves over time. In addition, for the purpose of personalizing the modeling of TSB dynamics across patients, we introduced two other parameters acting as time-shifting terms in the model: GA_i and Tc_i, denoting the GA at birth (in days) and the time correction factor, respectively.

Within the mathematical expression of the exponential decay, there are five parameters ($\boldsymbol{S}_i$), where GA_i is a known constant and the remaining four (A_i, B_i, C_i and Tc_i) are assigned during regression process of a robust non-linear least squared method [24, 25] in a patient-specific manner.

The regression (or curve fitting) process is equivalent to an optimization problem where we search $\boldsymbol{S}_i$ as the solution:

$$\sum_{i=1}^{n} \left(g(\boldsymbol{p}_i(t); \boldsymbol{S}_i) - \boldsymbol{TSB}_i\right)^2 \rightarrow \min_{\boldsymbol{S}_i} \tag{3.4}$$

To improve the identifiability of the problem, the parameters are carefully constrained with lower and upper bounds: $A \in [0, 200]$, $B \in [0, 1.5]$, $C \in [0, +\infty)$ and $Tc \in (-\infty, +\infty)$. Reasonable initial guesses for the parameters are provided based on prior knowledge and preliminary data analysis: $A_i = 100$, $B_i = 0.1$, $C_i = 50$ and $Tc_i = -GA_i$, where GA_i is the gestational age of patient i.

The robustness in the regression is achieved by assigning the samples with different weights, and this is configured by a loss function of a smooth approximation of L1 normalization controlled by a scaling parameter (refer to Section 2.3.2 for details on robust non-linear least squares). In addition, we have designed an adaptive robustness strategy to adjust the scaling parameter of the loss function, ensuring the efficiency and stability of the optimization process. This adaptive strategy assigns a unique factor to each subject based on a combination of the number of TSB measurements, the TSB value span, and its monotonicity, thereby facilitating a better adaptation of the proposed model to various physiological characteristics in different patients. The larger the value of the adaptive factor, the more robust the optimization process is, that is, the less it is affected by large deviations in the data.

Throughout the fitting process, parameters are iteratively adjusted within the specified boundaries to minimize the sum of squared residuals between the estimated and observed TSB values,

as stated in Equation 3.4. A parameter set ($\boldsymbol{S}_i$) that achieves the lowest sum of squared residuals is eventually identified as the model parameters for a given patient i. As a result of modeling, distinct parameter sets are defined (or fitted) for each patient individually, leading to a parameter collection ($\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{GA}$ and $\boldsymbol{Tc}$) that contains personalized parameters (A_i, B_i, C_i, GA_i and Tc_i) from every patient i.

3.2.4 Model analyses: from patient-specific models to clinical outcomes

With developed patient-specific TSB exponential decay models, further analyses were conducted.

First, the distribution of the model parameters was examined. Histograms were created to visualize the range and central tendencies of these parameters across the modeling population, providing insights into the variability, consistency and their contributions to overall model performance. The fitting error was evaluated by Root-mean-square error (RMSE) of the estimated and observed TSB levels for each patient:

$$\begin{aligned} \text{RMSE}_i &= \sqrt{\frac{1}{T_i}\sum_{t=1}^{T_i}(\hat{y}_i(t) - y_i(t))^2} \\ &= \sqrt{\frac{1}{T_i}(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i)^2} \\ &= \sqrt{\frac{1}{T_i}\boldsymbol{\epsilon}_i^2} \end{aligned} \tag{3.5}$$

where $\hat{y}_i$ and y_i denote the observed and estimated TSB levels for patient i at t (PNA) days, respectively; T_i is the number of TSB observations of patient i.

Next, we examined the association between the RMSE of the models and the occurrence of high-risk clinical events, such as NEC and CRP, aiming to determine whether the models were indicative of underlying clinical complications.

Finally, a **median model** was constructed using the median values of the parameters obtained from patient-specific models. A local sensitivity analysis was performed on this **median model** to investigate the impact of individual parameters on the model trajectory. This was implemented by varying each parameter one-at-a-time within its defined range and observing the effects on the model.

3.3 Results

3.3.1 Population and TSB measurements

The inclusion criteria of the study population are shown in the flowchart in Figure 3.5. A total of 373 eligible subjects contributed 2,208 total serum bilirubin (TSB) measurements during follow-ups. Of these, 85 infants with fewer than 4 measurements were excluded in the first stage. The remaining 288 infants, born between $24^{2/7}$ and $31^{6/7}$ weeks' gestation, were thus included, aggregating 2,011 TSB measurements in total. The data distribution of TSB measurements (µmol/L) against corresponding PNA (days) for this population is presented as beige scatter points in Figure 3.6, with histograms in beige above and to the right providing the distribution of PNA and TSB levels, respectively. In the histograms, the long tail in the PNA distribution suggests that most repeated measurements were taken in the early periods after birth. The positive skewness of the bilirubin distribution indicates that though some infants present elevated bilirubin levels, the general population tends to have progressively lower bilirubin levels over time.

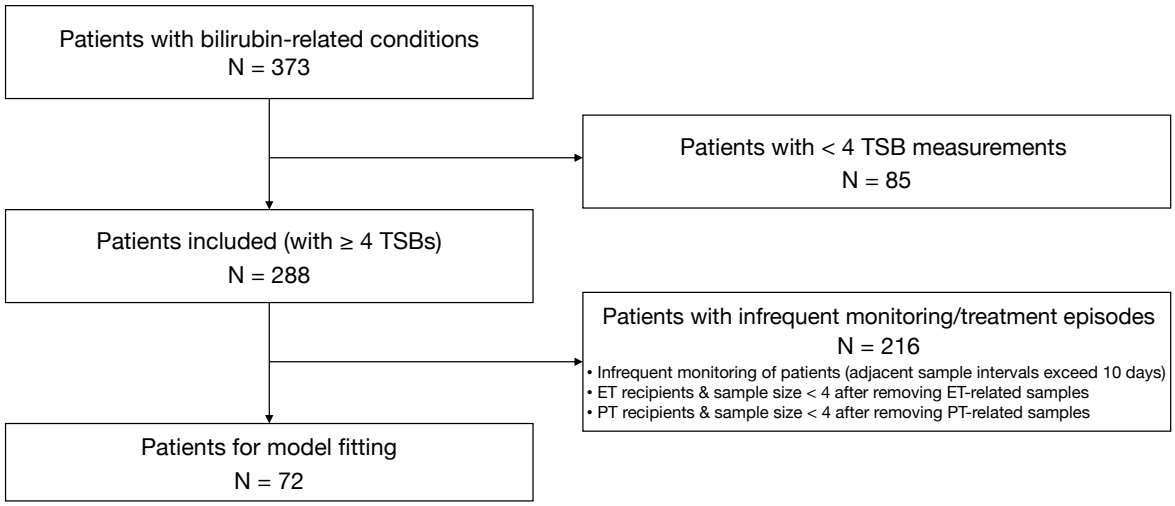

Figure 3.5: Study population inclusion criteria. *TSB*: total serum bilirubin. *ET*: exchange transfusion. *PT*: phototherapy.

Moreover, 32 infants were monitored infrequently, 3 were exchange transfusion recipients under NICE guidelines and 273 were phototherapy recipients as documented in the database. The data distribution for the 256 infants, after first excluding those who were not frequently monitored, is shown as blue data points in Figure 3.6, with histograms (in blue) of PNA and TSB levels on the top and right, respectively. The data presents a negative correlation (r=0.33) between PNA and corresponding TSB values. Of these three exceptions, some patients overlapped, so a total of 216 patients were excluded, leaving a subgroup of 72 patients for further analysis.

A total of 421 TSB samples from 72 patients were retained for modeling. The distribution of these data in the PNA-TSB plane is represented by orange points and histograms in Figure 3.7,

where all samples from the same patient are connected by orange lines. It also presents a negative correlation ($r = 0.26$) between PNA and corresponding TSB values. Statistically, the number of TSB measurements per infant is between 4 and 20 times with a median [IQR] of 5 [4; 6] samples. The initial bilirubin levels to be fit range from 128.6 µmol/L to 210 µmol/L with a mean (standard deviation [SD]) of 128.6 (31.1) µmol/L, and they were recorded at a median [IQR] PNA of 4.3 [2.0; 7.0] days. The last recorded TSB levels vary from 6 to 196 µmol/L with a mean (SD) of 70.8 (44.6) µmol/L, and they were acquired between 6.4 to 51.5 days after birth with a median [IQR] of 12.7 [9.7; 17.0] days.

Table 3.1 summarizes the general characteristics with descriptive statistics of the included 288 patients and the subgroup of 72 patients who were involved in patient-specific modeling. No significant difference was shown between the two populations.

3.3.2 General TSB decay models

Apart from the data distributions, both Figure 3.6 and Figure 3.7 also depict the results of general TSB decay models. Two general TSB decay models were respectively fitted to samples of two populations: N=288 and N=72 infants. In Figure 3.6, the exponential decay model achieved a marginally lower fitting error than the linear model (RMSE=45.82 µmol/L versus RMSE=44.81 µmol/L). In Figure 3.7, a subgroup of measurements, after excluding most of the unnatural values, two general models achieved similar trends and differences in fitting errors (RMSE=43.60 µmol/L versus RMSE=41.49 µmol/L) as in Figure 3.6.

Overall, bilirubin levels tend to decrease as postnatal age increases, but it is evident that the general models are not sufficient to accurately characterize the downward trend given significant large inter-individual variability. A great amount of data points notably deviated from the general curves, highlighting the necessity for more personalized modeling approaches.

3.3.3 Patient-specific models analyses

Patient-specific TSB exponential decay model

Patient-specific models were developed for 72 infants using selected TSB samples. For a given patient i, the function, as described in Equation 3.3, was personalized by finding the optimal set of parameters A_i, B_i, C_i and Tc_i that minimize the error between model output and the observations.

Histograms shown in Figure 3.8 depict the distributions of these parameters. Within predefined boundaries, the distribution of parameter A has a mean (SD) of 120.90 (27.54), comfortably within the limits (Figure 3.8a). Parameter B exhibits a median [IQR] of 0.11 [0.06; 0.21], clustering towards the lower limit (Figure 3.8b). Parameter C extends up to 171, but nearly three-quarters of the values are concentrated at the lower boundary of 0 (Figure 3.8c). Parameter Tc ranges from -264.7 to -51.4, with a mean (SD) of -194.9 (42.0) (Figure 3.8d). The mean (SD) RMSE of 72

Table 3.1: General characteristics of the included population and the subgroup used in patient-specific modeling.

	Characteristics	Included Patients *N*=288	Modeled Patients *N*=72	*P*-value
Initial conditions	Multiple pregnancy, n (%)	87 (30.2%)	15 (20.8%)	0.152
	Hypertension in pregnancy, n (%)	54 (18.8%)	18 (25.0%)	0.307
	Preterm labor, n (%)	192 (66.7%)	48 (66.7%)	1.000
	Chorioamnionitis, n (%)	19 (6.60%)	4 (5.56%)	0.957
	Corticosteroids, n (%)	268 (93.1%)	67 (93.1%)	1.000
	Delivery route, n (%)			0.506
	Vaginal delivery	121 (42.0%)	34 (47.2%)	
	C-section	167 (58.0%)	38 (52.8%)	
	GA at birth (weeks), mean (SD); median	28.1 (1.74); 28.1	27.8 (1.72); 27.8	0.218
	Birth weight (g), mean (SD); median	1082 (295); 1050	1009 (260); 960	0.051
	Birth weight Z-score[†], mean (SD); median	−0.03 (0.80); 0.05	−0.10 (0.87); 0.14	0.767
	Gender (male), n (%)	160 (55.6%)	36 (50.0%)	0.475
	Apgar (1-min score), median [IQR]	6.00 [3.00; 8.00]	6.00 [2.00; 8.00]	0.715
	Intubation at birth, n (%)	62 (21.5%)	11 (15.3%)	0.310
	PDA on PNA = 4 days, n (%)	139 (48.3%)	40 (55.6%)	0.330
Outcomes	Neurologic impairment, n (%)	76 (26.4%)	17 (23.6%)	0.741
	Respiratory support stopped before 34 PMA weeks, n (%)	77 (26.7%)	15 (20.8%)	0.381
	Death, n (%)	16 (5.56%)	4 (5.56%)	1.000
	PNA at death (days), median [IQR]	35.0 [19.5; 69.6]	43.3 [20.0; 98.4]	0.813
	PMA at death (weeks), median [IQR]	33.2 [28.7; 36.6]	31.9 [27.8; 40.2]	0.813
	Interruption of follow-up, n (%)	2 (0.69%)	0 (0.00%)	1.000
	Length of follow-up (days), mean (SD); median	30.0 (16.1); 27.2	32.9 (15.0); 28.3	0.106
	Phototherapy, n (%)	273 (94.8%)	62 (86.1%)	0.020

[†]Z-scored birth weight based on gestational age according to Fenton's 2013 preterm growth chart [21].

Features are divided into initial conditions and outcomes according to the acquisition time before or after PNA = 4 days.

The statistics were reported as **counts (percentage)** for categorical variables and as **mean (SD); median** or **median [IQR]** for continuous variables, as appropriate. Categorical variables were compared using the chi-square test and continuous variables using the Mann-Whitney U test, as appropriate.

The significance level was set to 0.05 and adjusted by Bonferroni correction, thus in this case, p-value <0.0025 was considered significant.

GA: gestational age. *PDA*: patent ductus arteriosus. *PNA*: postnatal age. *PMA*: postmenstrual age. *SD*: standard deviation. *IQR*: interquartile range.

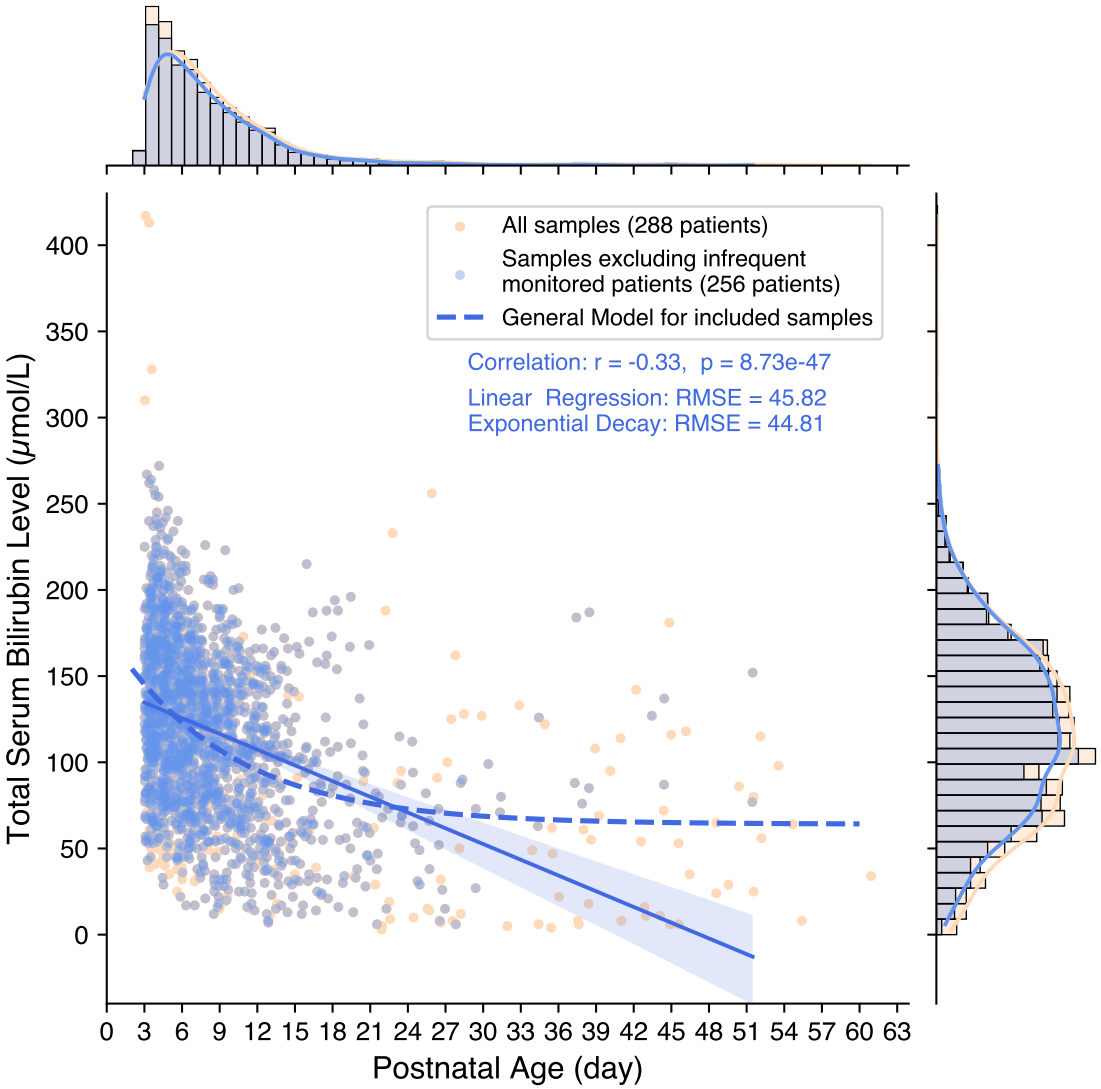

Figure 3.6: Total serum bilirubin (TSB) levels in µmol/L relative to postnatal age (PNA) in days for the included population (N=288) and a subgroup after excluding infrequently monitored patients (N=256). General models of linear regression (solid blue line) and exponential decay (dashed blue curve). The correlation between TSB in µmol/L and PNA in days was assessed using Pearson's correlation coefficient.

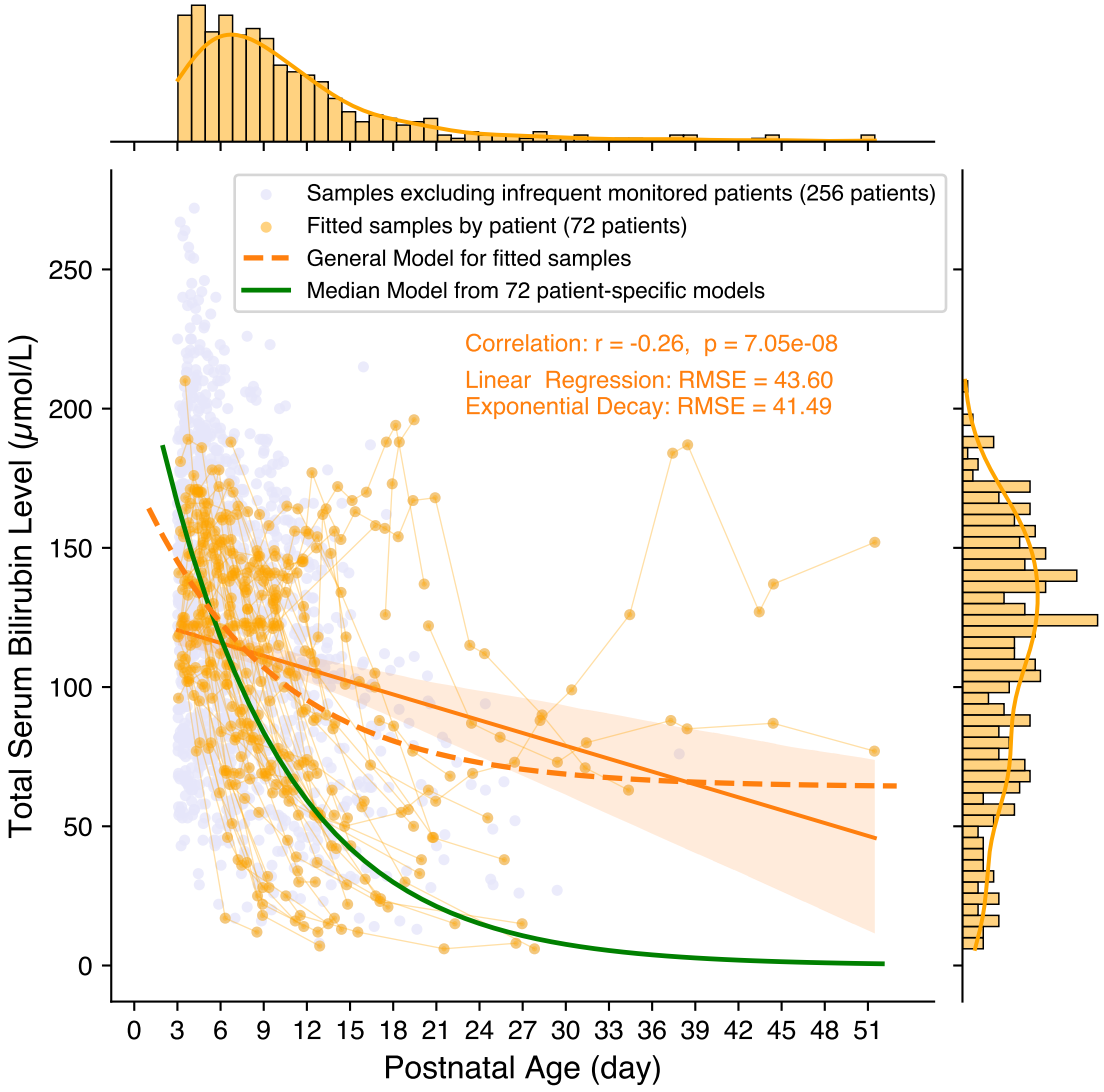

Figure 3.7: TSB measurements in μmol/L relative to PNA in days for the fitted population (N=72), in which all samples from the same patient are connected by light orange lines. General models of linear regression (solid orange line) and exponential decay (dashed orange curve). The **median model** (solid green curve) from patient-specific models. The correlation between TSB in μmol/L and TSB in days was assessed using Pearson's correlation coefficient.

models is 10.62 (7.69) µmol/L, with a median [IQR] of 8.74 [4.89; 14.25] µmol/L, as shown in Figure 3.8e.

Moreover, Figure 3.8f presents both the counts and values of elevated C-reactive protein (CRP) levels (defined by CRP >5 mg/L) across different RMSE values, solely considering CRP results obtained post-phototherapy and before the last TSB measurements. Despite the limited sample size, a clear trend emerges: lower RMSE values correspond to fewer instances and lower maximum values of elevated CRP levels, whereas higher RMSE values are associated with an increase in both occurrences and higher maximum values of elevated CRP.

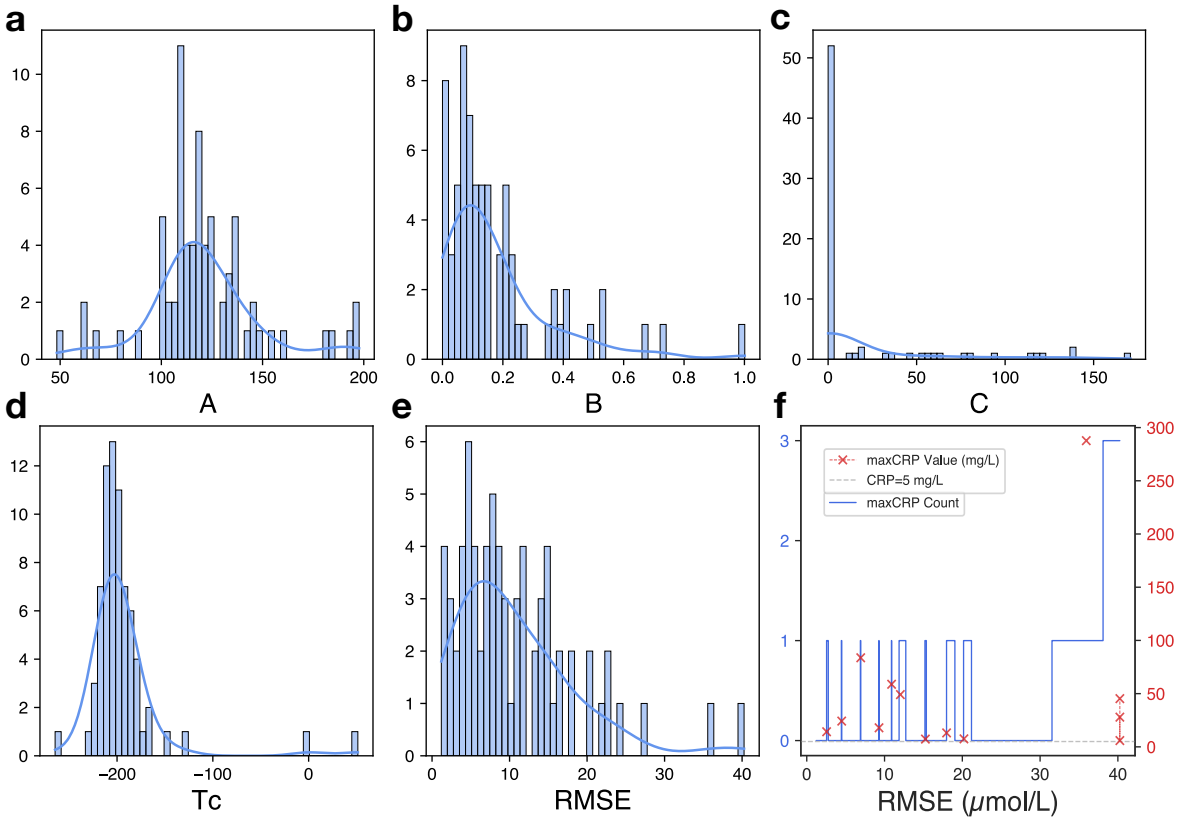

Figure 3.8: Histograms of parameters and fitting errors of patient-specific models among 72 patients. (**a**) Parameter *A*. (**b**) Parameter *B*. (**c**) Parameter *C*. (**d**) Parameter *Tc*. (**e**) RMSE. (**f**) Relationship between RMSE and counts of CRP greater than 5 mg/L (blue step plot) and maximum CRP values (red scatter plot), respectively.

Figure 3.9 illustrates 9 well-fitted instances of patient-specific models exhibiting various curve morphologies ordered by increasing decay rates. Of these, models for Patients 1173, 1178, and 1243 shown in Figure 3.9c, Figure 3.9g and Figure 3.9h were marked with significantly high CRP levels between the two samples when their bilirubin levels were not frequently monitored, whereas Patients 1188 (Figure 3.9e) and 1162 (Figure 3.9i) experienced elevated CRP days after the cessation of TSB monitoring. All patient-specific models of 72 patients are presented in Appendix Figure A.1 to Figure A.8.

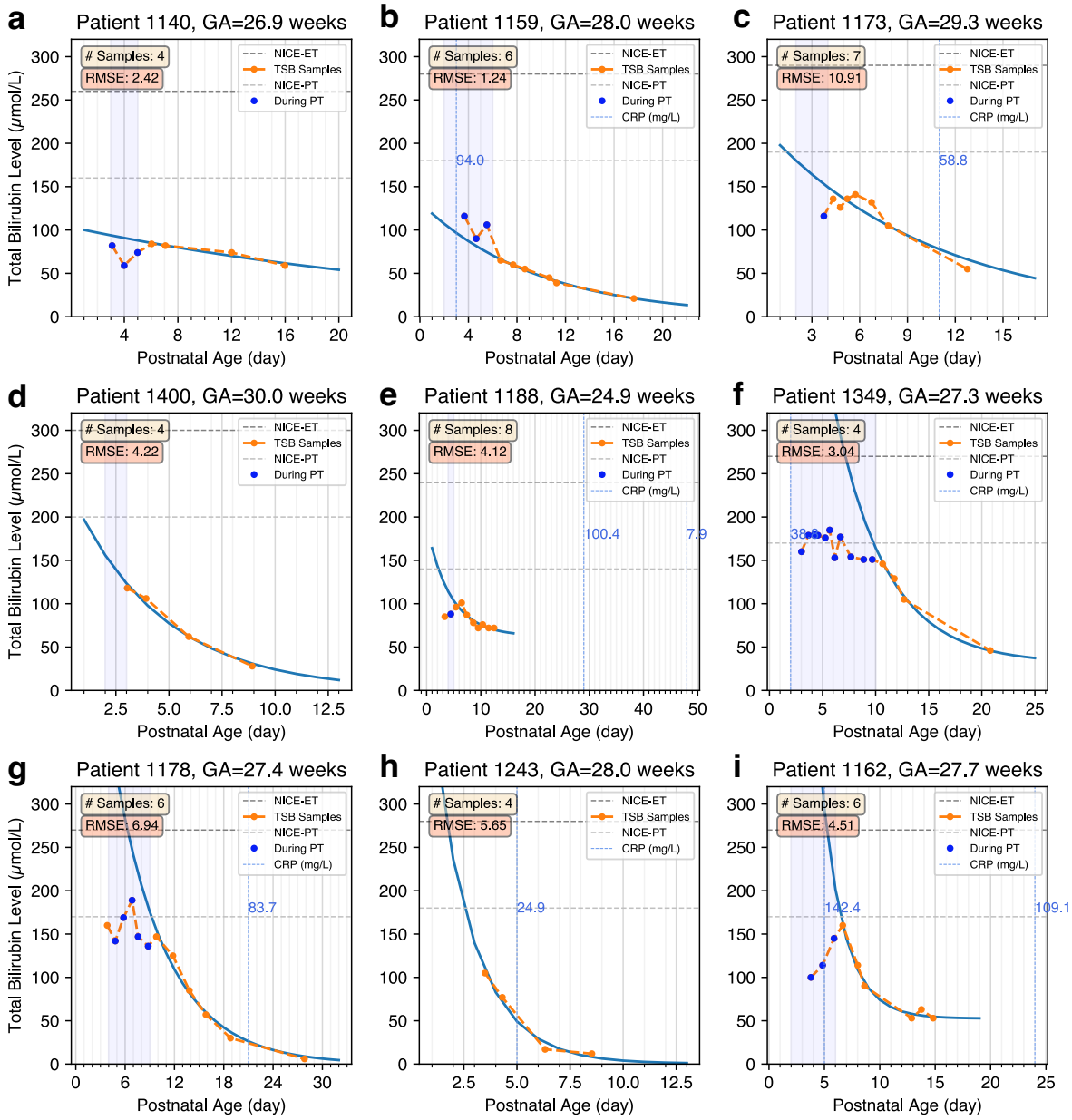

Figure 3.9: Nine representative instances of patient-specific models (blue solid curves) exhibiting various curve morphologies on the PNA-TSB plane, ordered by increasing rates of decay. (**a**) Patient 1140. (**b**) Patient 1159. (**c**) Patient 1173. (**d**) Patient 1400. (**e**) Patient 1188. (**f**) Patient 1349. (**g**) Patient 1178. (**h**) Patient 1243. (**i**) Patient 1162. The x-axes are Postnatal age (PNA) in days; the y-axes are Total serum bilirubin (TSB) in μmol/L. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNA. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

From patient-specific models to clinical outcomes

Figure 3.10 presents instances of great variability in terms of TSB trends, modeling errors, and clinical outcomes. Patient 1102 (Figure 3.10a) is an extremely preterm infant born at $25^{6/7}$ weeks of gestation with a birth weight of 620 grams. Both 1-minute and 5-minute Apgar scores were 1, indicating a critically low health status and poor adaptation. This baby was diagnosed with NEC on day 31 of life, five days after the last bilirubin measurement. The remaining five models depict infants who experienced recurrent hyperbilirubinemia after 2 weeks of life, with TSB levels exceeding the phototherapy thresholds recommended by the NICE guidelines. These models have the highest fitting errors evaluated by RMSE: four of them have the top four RMSE values and one model ranks seventh in RMSE among the 72 patient-specific models.

The TSB of Patient 1336 (Figure 3.10b) fluctuated around the phototherapy threshold, peaking at 194 µmol/L around day 18. The patient had grade *i*-a enterocolitis from days 9 to 16, without inflammation or elevated CRP documented, corresponding to a rising TSB episode.

Patient 1317 (Figure 3.10c) shows a notable upward trend in TSB levels post-phototherapy, peaking on day 14. This infant had a persistent Patent ductus arteriosus (PDA) that worsened progressively during the first two weeks and ended with surgical intervention on day 17. TSB monitoring ceased afterward, but the infant developed a late-onset infection on day 19 and died of septic shock the same day.

Similarly, Patient 1271 (Figure 3.10d) was born at $25^{2/7}$ weeks of gestation with 755 grams of birth weight. A significant PDA was developed from day 2, which was surgically closed on day 15. Besides, a localized grade 4 Intraventricular hemorrhage (IVH) on day 3 which resolved progressively on subsequent ultrasound scans.

Figure 3.10e presents a complex case of Patient 1412, whose bilirubin levels dropped below 100 µmol/L post-phototherapy by day 11. However, a TSB rebound occurred on day 13, coinciding with an extreme CRP level of 287.7 mg/L and a NEC diagnosis. Then a second hyperbilirubinemia rebound was observed around day 38 of life.

Patient 1373 (Figure 3.10f) underwent a notable decline in TSB levels after day 5, followed by a drastic rebound, with TSB fluctuating below the PT threshold. The infant developed cholestasis starting on day 5 (conjugated bilirubin 57 µmol/L), reaching a maximum on day 12 (conjugated bilirubin 173 µmol/L). Etiological investigations remained inconclusive, and the cholestasis eventually resolved. Subsequently, the infant developed grade i-a enterocolitis on day 17, with elevated CRP (45.3 mg/L) and a positive blood culture.

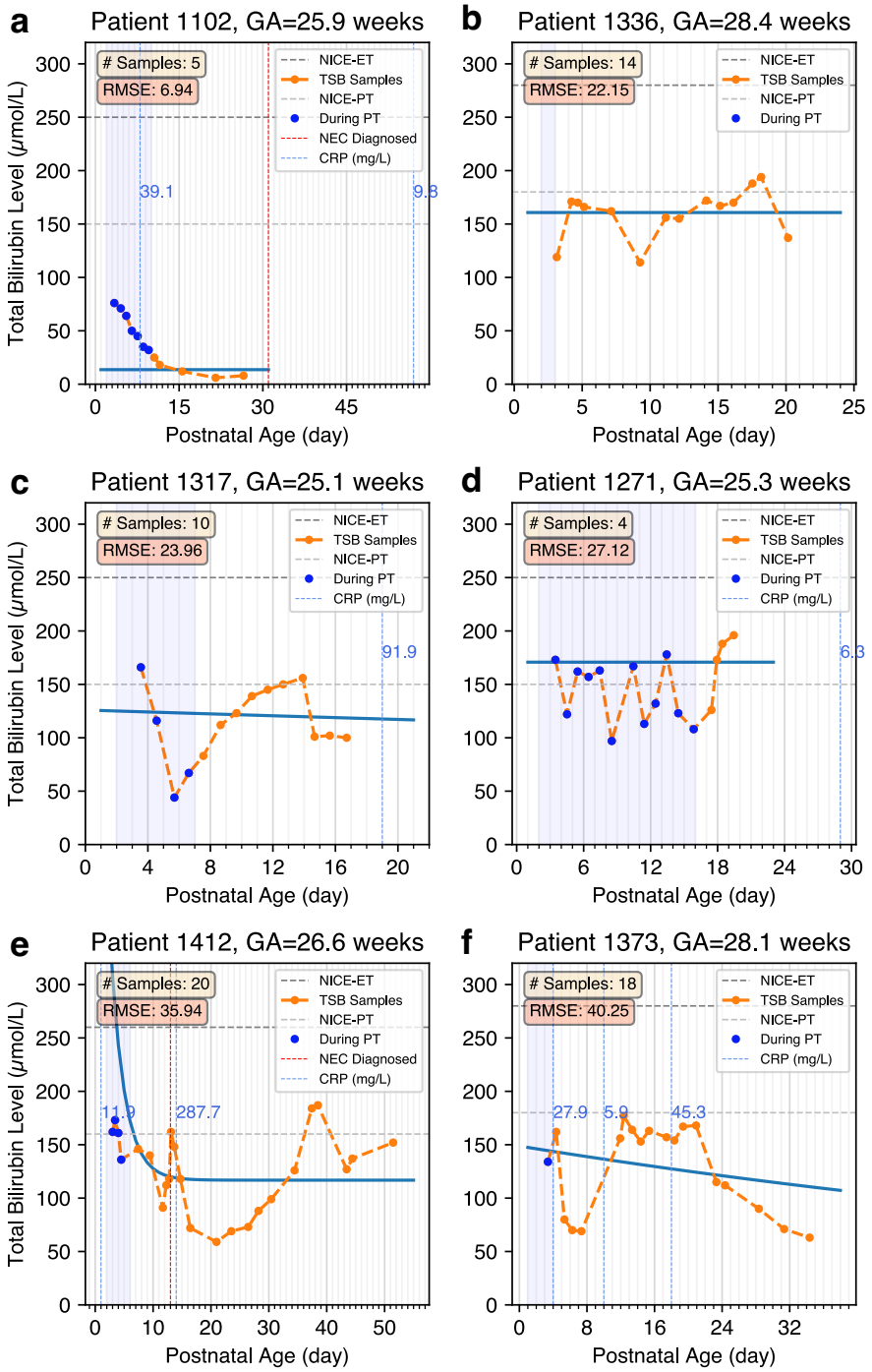

Figure 3.10: Patient-specific models with variability in terms of bilirubin trends, modeling errors (RMSE) and clinical outcomes. (**a**) Patient 1102. (**b**) Patient 1336. (**c**) Patient 1317. (**d**) Patient 1271. (**e**) Patient 1412. (**f**) Patient 1373. The x-axes are Postnatal age (PNA) in days; the y-axes are Total serum bilirubin (TSB) in µmol/L. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue vertical dashed lines are CRP values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey). Red vertical dashed lines mark NEC diagnosis.
ET: exchange transfusion. *PT*: phototherapy. *CRP*: C-reactive protein. *NEC*: necrotizing enterocolitis.

Local sensitivity analysis of patient-specific model parameters

The **median model**, depicted by the green curve in Figure 3.7, is constructed using the median values from the parameter collection derived from 72 patient-specific models, and it can be instantiated as:

$$\textbf{Median model} : 118.3388 \times exp\left(-0.1142\left[\boldsymbol{p}(t) + 196 - 201.9616\right]\right) + 0 \tag{3.6}$$

By systematically varying each parameter within a certain range, we were able to observe and qualify the resulting changes in the patient-specific model's behavior. The results reveal distinct patterns in their effects.

In Figure 3.11, a series of curves represent models fitted with varying values of parameter $\boldsymbol{A}$, ranging from 0 to 400. As the values of $\boldsymbol{A}$ increases, the curve shifts rightward and upward. Figure 3.12 shows the effect of the parameter $\boldsymbol{B}$, the decay factor that controls the rate of decrease in the exponential function. The decay rate sharply rises as parameter $\boldsymbol{B}$ increases from 0 to 0.3. Figure 3.13 demonstrates how the curve is elevated as parameter $\boldsymbol{C}$ varies from -100 to 200. Lastly, Figure 3.14 depicts the local sensitivity analysis of parameter $\boldsymbol{Tc}$, which controls the time shift within the support range of -215 to -150 days. The curve shifts further rightward as $\boldsymbol{Tc}$ becomes more negative. The influence of $\boldsymbol{GA}$ in the model behaves similarly to $\boldsymbol{Tc}$, as both parameters occupy the same position in the proposed Equation 3.3.

3.4 Discussion

In this study, we proposed and quantitatively evaluated a patient-specific exponential decay model for characterizing the dynamics of Total serum bilirubin (TSB) levels with age (PNA) in a population of very preterm infants during the first weeks of life, as formalized in Equation 3.3. To the best of our knowledge, this is the first study focusing on the natural history of TSB levels over such a long-term course in preterm infants born at 24-32 weeks of gestation.

The downward trend captured by the proposed exponential decay model is strongly associated with bilirubin metabolism. Preterm infants typically exhibit high bilirubin levels after birth, which is commonly referred to as physiological neonatal jaundice. Interventions such as phototherapy are often administered to manage TSB levels. With treatment, TSB levels usually fall below critical thresholds and start to drop at a high rate of decline. As the infants age, their metabolic systems become more developed and refined, progressively achieving an equilibrium between bilirubin production and elimination. This process is characterized by a decelerating rate of decay and a relatively smooth tail, as depicted in the median evolutionary model (green curve in Figure 3.7). We propose that the evolution of these physiological phenomena can be effectively captured by specific model parameters.

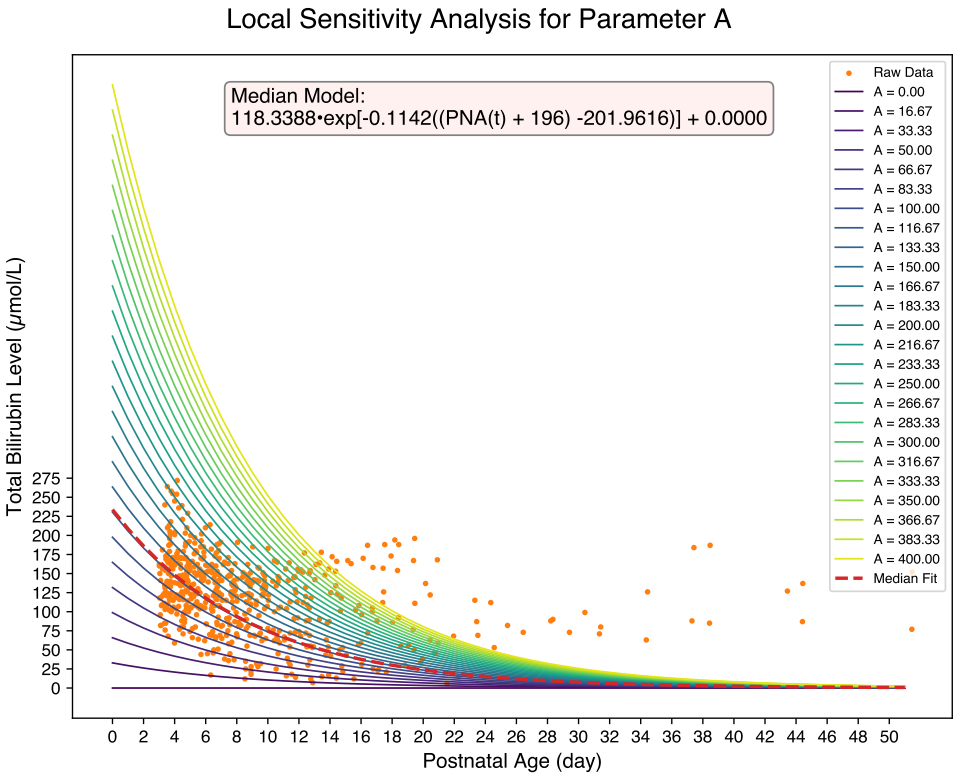

Figure 3.11: Local sensitivity analysis of parameters A on the **median model**.

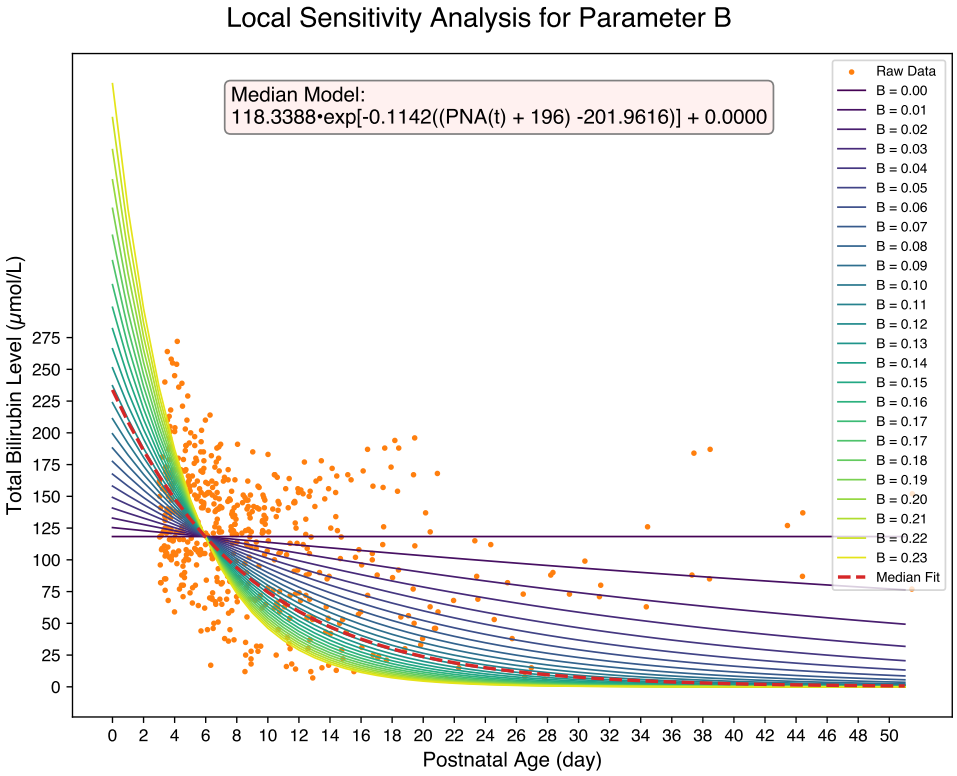

Figure 3.12: Local sensitivity analysis of parameters B on the **median model**.

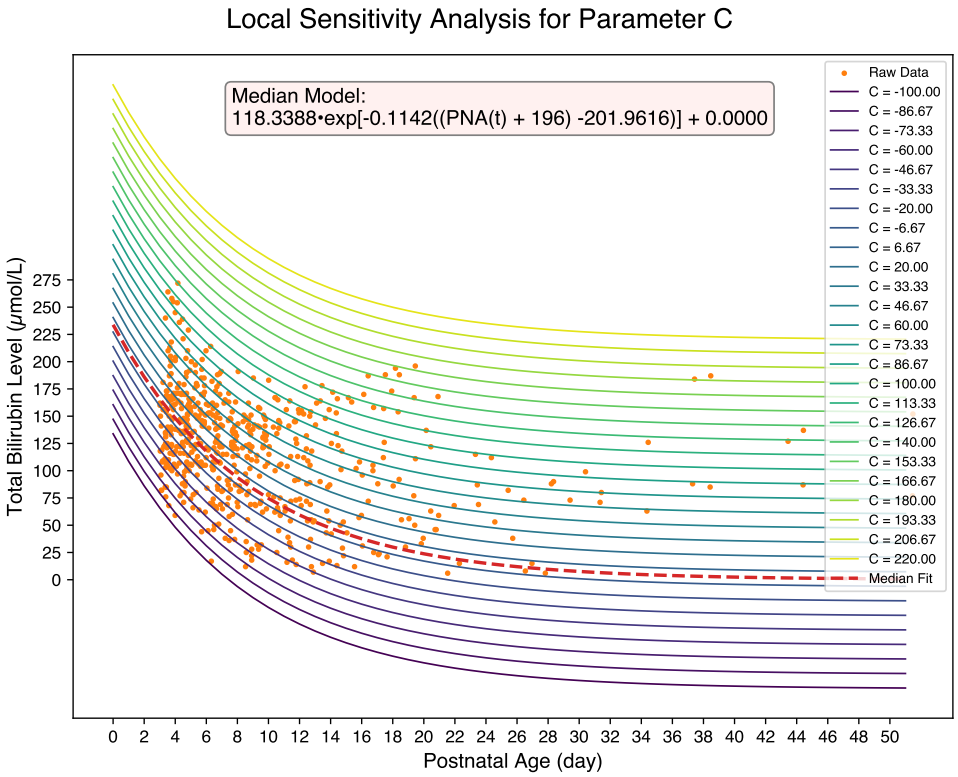

Figure 3.13: Local sensitivity analysis of parameters ***C*** on the **median model**.

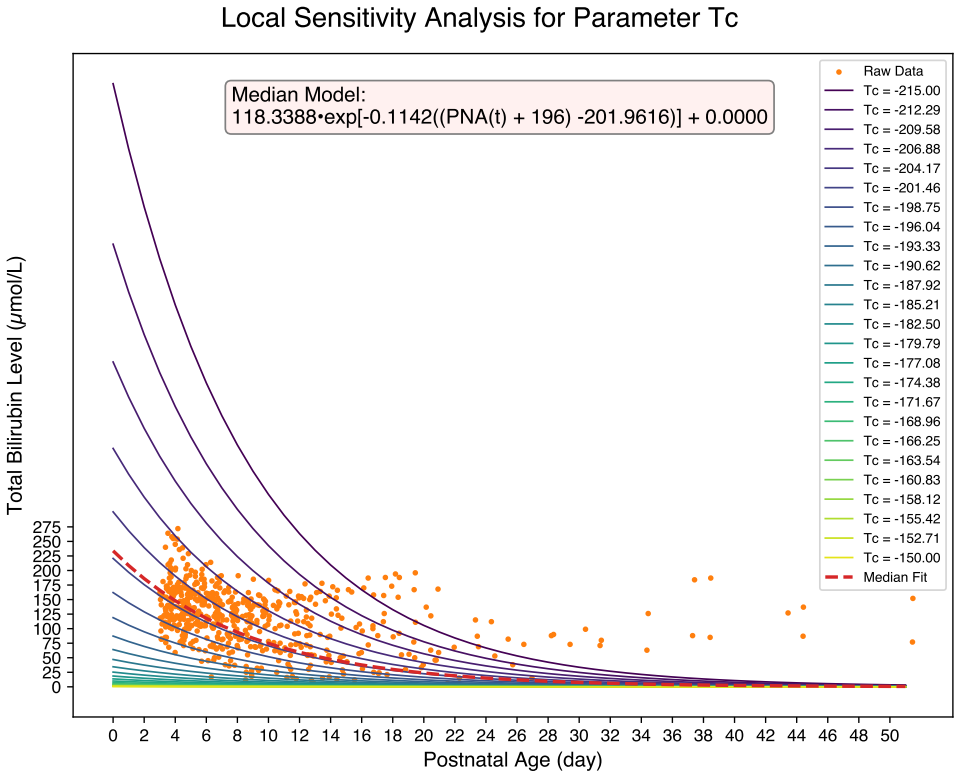

Figure 3.14: Local sensitivity analysis of parameters ***Tc*** on the **median model**.

The local sensitivity analysis of the **median model** elucidates the specific effects of each parameter on the model, detailing how changes in parameter values sculpt the model. Parameter $\boldsymbol{A}$ in the proposed model controls the initial concentration of bilirubin at the start of the modeled time period. It is linked to which is influenced by bilirubin production (e.g., red blood cell breakdown) and elimination (e.g., the ability of the liver to process bilirubin), as shown in the first steps of Figure 1.3. Parameter $\boldsymbol{B}$ represents the rate of enzymatic and transport processes involved in bilirubin metabolism. It characterizes the rate at which bilirubin levels decline over time, reflecting the efficiency of key clearance mechanisms, including hepatic conjugation and biliary excretion. Parameter $\boldsymbol{C}$ represents the asymptotic bilirubin concentration as time progresses, indicating a residual or baseline level, and it corresponds to the minimum bilirubin level that the infant can achieve, which depends on metabolic and excretion capabilities. Parameter $\boldsymbol{GA}$ acts as a constant representing the gestational age at birth (in days), affecting the maturity of the liver and other metabolic systems, which in turn influence bilirubin metabolism and clearance. Parameter $\boldsymbol{Tc}$ adjusts for individual variations in time-dependent bilirubin metabolism, used to capture personalized temporal shifts in bilirubin dynamics such as delayed liver function maturation, allowing the model to account for inter-individual variability in metabolic response. Unlike the other three parameters that function independently, we explicitly decompose the time-shift term in the proposed model into two components: time correction factor ($\boldsymbol{Tc}$) and gestational age ($\boldsymbol{GA}$) at birth. This decomposition enhances the explainability.

GA at birth is a key indicator of neonatal maturity, and its crucial impact has been consistently evidenced in the literature. It is reported [26] that including GA in their assessment model for estimating the risk of significant hyperbilirubinemia in infants significantly improved predictive accuracy. Several reports have also noted that GA (and BW) plays a decisive role in the initiation and duration of phototherapy [27, 28]. The importance of GA is further corroborated by the NICE [1] and most national guidelines, which strictly differentiate treatment thresholds for phototherapy and exchange transfusion in neonates based on gestational age.

In addition to GA, we explicitly incorporated another term, parameter $\boldsymbol{Tc}$, as part of the "time-shifter" in the model to account for varying maturity levels among neonates of the same gestational age. The results show that $\boldsymbol{Tc}$ values are all negative, this is because the values of $\boldsymbol{GA}$ are typically in hundreds of days, requiring $\boldsymbol{Tc}$ to act as a compensatory factor to bring the integrated time-shift term into the normal range.

By formulating a patient-specific function that describes postnatal age-based TSB levels as an exponential decay model, this study provides a more comprehensive description of the postnatal bilirubin decline in preterm infants with a gestational age of 24-32 weeks. On the one hand, the patient-specific model and its associated parameters reflect the similarities and variations of the natural development of bilirubin metabolism among the considered population. On the other hand, we demonstrated the predictive ability of the proposed model by monitoring deviations of actual bilirubin levels from the fitted natural bilirubin trends, with RMSE serving as one indicator

to quantify these deviations. We illustrated this by annotating elevated CRP levels on the fitted models as shown in Figure 3.9 and Figure 3.10. The discrepancies between measured TSB levels and the expected TSB trend identified by the proposed model could offer a valuable tool for disease detection, phenotyping, and prediction in the NICU. If a newly measured TSB level deviates from the expected decay, it could alert caregivers and prompt them to check for possible comorbidities such as infections, NEC, or other complications in the infant. A potential application could be a clinical decision support system based on interpretable models for optimizing NICU monitoring and detecting high-risk events.

Additionally, as shown in Figure 3.9 and Figure 3.10, the lack of TSB measurements around elevated CRP levels exposes us to the risk of overlooking critical information. Therefore, it may be necessary to increase both the frequency and the duration of TSB monitoring. This would facilitate better observation of bilirubin dynamics and suspicious complications reflected by unnatural fluctuations in bilirubin. In clinical practice, TSB can be obtained through micro-sampling during other required biological monitoring without additional blood spoliation.

Nevertheless, this study has several limitations. First, the model-fitting phase excluded patients with fewer than 4 TSB measurements. We observed significant differences in 5 of the 13 initial conditions and 4 of the 8 outcomes between the included and excluded populations, as detailed in Appendix Table A.1. This exclusion might introduce selection bias and limit the generalization of the proposed model. Second, due to the limited sample size, we could only perform our analysis of the association between the proposed model and clinical events through qualitative assessments of a small number of cases. Future research could be dedicated to more systematic and extensive data collection and focus on quantitative analyses based on a richer database to evaluate the utility and performance of the model and the feasibility of using bilirubin levels as a potential indicator of high-risk clinical events in the NICU. Lastly, the CARESS-Premi clinical protocol only enrolled patient data from three days after birth. Consequently, the TSB evolution models proposed in this study lack information for the first three days. However, there is a large literature on the hourly trends of TSB in preterm infants within the first 72 or 96 hours after birth [13, 29], and what our research presented plays a role as an extension and supplement to these previous studies, i.e., when combined, a comprehensive pattern of TSB development from birth to the several weeks postnatally could be portrayed.

3.5 Conclusion

In this chapter, we developed and validated a patient-specific exponential decay model to characterize the natural and long-course dynamics of total serum bilirubin concentrations in preterm infants born between 24 to 32 weeks of gestation. Our approach originally leverages a deterministic mathematical model to describe bilirubin kinetics over the initial weeks of like, providing an unprecedented long-term view of TSB evolution in this vulnerable population.

Through personalized model parameter fitting, we obtained 72 patient-specific models, optimized by minimizing the error between measured TSB and model output using an adaptive robust least-squared method. The proposed model demonstrated its effectiveness and capability to closely track observed TSB levels during extended neonatal periods, with an RMSE ranging from 1.20 µmol/L to 40.25 µmol/L, with a median [IQR] of 8.74 [4.89; 14.25] µmol/L. The natural course of TSB evolution follows the proposed exponential decay trend, with variations in the parameters between models reflecting individual differences in bilirubin metabolism and developmental patterns in the studied population.

Furthermore, when the bilirubin evolutionary trend of a given patient diverges from the expected decay pattern, as indicated by an increased RMSE, it might suggest the occurrence of high-risk clinical events such as necrotizing enterocolitis and elevated C-reactive protein levels. This association indicates that the model's capabilities extend beyond mere descriptive analytics and may serve as a new digital tool for the early detection of relevant comorbidities. Furthermore, the analysis of the obtained patient-specific model parameters might be useful for clinical phenotyping, in order to better characterize this population and improve diagnostic and therapeutic strategies.

Moving forward, the integration of this model into a model-based clinical decision support system holds significant promise for enhancing NICU monitoring optimization and facilitating high-risk event detection. This advancement could greatly improve the management and outcomes for preterm infants.

In the next chapter, this proposed exponential decay model, with physiological insights, will be embedded and evaluated in knowledge-based estimators, facilitating non-invasive bilirubin estimations in the context of longitudinal clinical data in the NICU setting.

BIBLIOGRAPHY

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

[2] A. R. Kemper, T. B. Newman, J. L. Slaughter, M. J. Maisels, J. F. Watchko, S. M. Downs, R. W. Grout, D. G. Bundy, A. R. Stark, D. L. Bogen et al., "Clinical practice guideline revision: management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation," Pediatrics, vol. 150, no. 3, 2022.

[3] C. J. Wusthoff and I. M. Loe, "Impact of bilirubin-induced neurologic dysfunction on neurodevelopmental outcomes," Seminars in Fetal and Neonatal Medicine, vol. 20, no. 1, pp. 52–57, feb 2015.

[4] J. F. Watchko, "Bilirubin-induced neurotoxicity in the preterm neonate," Clinics in Perinatology, vol. 43, no. 2, pp. 297–311, 2016.

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

[6] V. K. Bhutani, L. Johnson, and E. M. Sivieri, "Predictive ability of a predischarge hour-specific serum bilirubin for subsequent significant hyperbilirubinemia in healthy term and near-term newborns," vol. 103, no. 1, pp. 6–14, 1 1999.

[7] D. De Luca, G. L. Jackson, A. Tridente, V. P. Carnielli, and W. D. Engle, "Transcutaneous bilirubin nomograms: a systematic review of population differences and analysis of bilirubin kinetics," vol. 163, no. 11, 11 2009.

[8] M. Kaplan and M. J. Maisels, "Natural history of early neonatal bilirubinemia: A global perspective," vol. 41, no. 4, pp. 873–878, 4 2021.

[9] American Academy of Pediatrics Subcommittee on Hyperbilirubinemia, "Management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation," Pediatrics, vol. 114, no. 1, pp. 297–316, 2004.

[10] S. Onishi, N. Kawade, S. Itoh, K. Isobe, and S. Sugiyama, "Postnatal development of uridine diphosphate glucuronyltransferase activity towards bilirubin and 2-aminophenol in human liver," Biochemical Journal, vol. 184, no. 3, pp. 705–707, 1979.

[11] M. J. Maisels, V. K. Bhutani, D. Bogen, T. B. Newman, A. R. Stark, and J. F. Watchko, "Hyperbilirubinemia in the newborn infant $\geq$35 weeks' gestation: an update with clarifications," Pediatrics, vol. 124, no. 4, pp. 1193–1198, 2009.

[12] M. Maisels, J. Watchko, V. Bhutani, and D. Stevenson, "An approach to the management of hyperbilirubinemia in the preterm infant less than 35 weeks of gestation," Journal of perinatology, vol. 32, no. 9, pp. 660–664, 2012.

[13] S. Hahn, C. Bührer, G. Schmalisch, B. Metze, and M. Berns, "Rate of rise of total serum bilirubin in very low birth weight preterm infants," vol. 87, no. 6, pp. 1039–1044, 5 2020.

[14] T. Jegathesan, J. G. Ray, V. K. Bhutani, C. D. G. Keown-Stoneman, D. M. Campbell, V. Shah, H. Berger, R. Z. Hayeems, M. Sgro, and NeoHBC, "Hour-specific total serum bilirubin percentiles for infants born at 29–35 weeks' gestation," Neonatology, vol. 118, no. 6, pp. 710–719, 2021.

[15] T. Jegathesan, J. G. Ray, C. D. G. Keown-Stoneman, D. M. Campbell, V. Shah, H. Berger, R. Z. Hayeems, and M. Sgro, "Pre-phototherapy total serum bilirubin levels in extremely preterm infants," Pediatric Research, vol. 93, no. 1, pp. 226–232, 2023.

[16] D. Mukherjee, M. Coffey, and M. J. Maisels, "Frequency and duration of phototherapy in preterm infants <35 weeks gestation," Journal of Perinatology, vol. 38, no. 9, pp. 1246–1251, 2018.

[17] M. Kaplan, E. Kaplan, C. Hammerman, N. Algur, R. Bromiker, M. S. Schimmel, and A. I. Eidelman, "Post-phototherapy neonatal bilirubin rebound: a potential cause of significant hyperbilirubinaemia," Archives of disease in childhood, vol. 91, no. 1, pp. 31–34, 2006.

[18] N. Kawade and S. Onishi, "The prenatal and postnatal development of UDP-glucuronyltransferase activity towards bilirubin and the effect of premature birth on this activity in the human liver," The Biochemical Journal, vol. 196, no. 1, pp. 257–260, 1981.

[19] S. Itoh, H. Okada, K. Koyano, S. Nakamura, Y. Konishi, T. Iwase, and T. Kusaka, "Fetal and neonatal bilirubin metabolism," Frontiers in Pediatrics, vol. 10, 2023.

[20] M. J. Bell, J. L. Ternberg, R. D. Feigin, J. P. Keating, R. Marshall, L. Barton, and T. Brotherton, "Neonatal necrotizing enterocolitis: Therapeutic decisions based upon clinical staging," Annals of Surgery, vol. 187, no. 1, p. 1, 1978.

[21] T. R. Fenton and J. H. Kim, "A systematic review and meta-analysis to revise the Fenton growth chart for preterm infants," BMC Pediatrics, vol. 13, no. 1, p. 59, 2013.

[22] S. Mitra and J. Rennie, "Neonatal jaundice: Aetiology, diagnosis and treatment," British Journal of Hospital Medicine, vol. 78, no. 12, pp. 699–704, 2017.

[23] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in 9th Python in Science Conference, 2010.

[24] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas,

D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," Nature Methods, vol. 17, pp. 261–272, 2020.

[25] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings. Springer, 2000, pp. 298–372.

[26] R. Keren, X. Luan, S. Friedman, S. Saddlemire, A. Cnaan, and V. K. Bhutani, "A comparison of alternative risk-assessment strategies for predicting significant neonatal hyperbilirubinemia in term and near-term infants," Pediatrics, vol. 121, no. 1, pp. e170–e179, 2008.

[27] K. Mreihil, J. Š. Benth, H. J. Stensvold, B. Nakstad, T. W. R. Hansen, the Norwegian NICU Phototherapy Study Group, and the Norwegian Neonatal Network, "Phototherapy is commonly used for neonatal jaundice but greater control is needed to avoid toxicity in the most vulnerable infants," Acta Paediatrica, vol. 107, no. 4, pp. 611–619, 2018.

[28] D. Mukherjee, M. Coffey, and M. J. Maisels, "Frequency and duration of phototherapy in preterm infants < 35 weeks gestation," Journal of Perinatology, vol. 38, no. 9, pp. 1246–1251, 2018.

[29] T. Jegathesan, D. M. Campbell, J. G. Ray, V. Shah, H. Berger, R. Z. Hayeems, M. Sgro, and for the NeoHBC, "Transcutaneous versus total serum bilirubin measurements in preterm infants," vol. 118, no. 4, pp. 443–453, 2021.

CHAPTER

4 Bilirubin Estimation in Preterm Infants based on Modified Mixed-Effects Random Forests

This chapter is dedicated to another aspect of Neonatal hyperbilirubinemia (NHB) management, developing and more about evaluating the effectiveness of novel, knowledge-based, non-invasive methods for estimating bilirubin levels. In the context of longitudinal clinical data, approaches that leverage advanced monitoring resources and massive monitoring signals available in the NICU appear as promising tools for improving the care of neonates who suffer from significant hyperbilirubinemia. The exponential decay model proposed in Chapter 3 is also incorporated in this chapter to offer additional physiological insights that may empower the performance in bilirubin estimations.

Therefore, in this study, we propose novel TSB estimators based on Mixed-Effects Random Forest (MERF) that incorporate specific physiological insights as additional mixed-effects terms. The performance gains of these models when compared to standard Random Forest models are then assessed.

4.1 Introduction

Hyperbilirubinemia, characterized by elevated levels of bilirubin in the blood, is particularly prevalent and significant in preterm infants due to their greater degree of immaturity, underdeveloped hepatic function and higher susceptibility to complications [1–3]. If not managed properly, high bilirubin levels can induce irreversible neurotoxicity, leading to severe neurological damage such as kernicterus [4–6], and it is associated with adverse neurodevelopmental outcomes as reviewed in Section 1.3.1.

Regular Total serum bilirubin (TSB) monitoring is crucial for timely interventions such as pho-

totherapy and exchange transfusion. However, this typically involves frequent blood sampling, which is invasive and painful for these tiniest babies and repetitive procedures can lead to complications such as anemia, infection and osteomyelitis [7]. Non-invasive methods for estimating bilirubin levels are thus gaining more attention. These methods should be developed with the aim of optimizing the timing of blood sampling and ultimately minimizing blood exploitation and associated risks. Transcutaneous bilirubin (TcB) measurements have been widely used as a non-invasive alternative despite slight discrepancies compared to serum levels [3, 8].

On the other hand, fortunately, modern Neonatal intensive care units (NICU) are equipped with advanced monitoring resources, generating continuous data from admitted preterm babies, particularly ECG recordings. This continuous monitoring presents a promising opportunity to develop new tools in this subject. Research has shown that hyperbilirubinemia influences the activity of the Autonomic nervous system (ANS), especially through parasympathetic predominance, manifested by alterations in Heart rate variability (HRV) in jaundiced term and preterm neonates [9–12]. 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 models.

The selection of the appropriate machine learning algorithm is a challenge in this field, characterized by longitudinal, time-dependent and noisy data. Indeed, longitudinal data exhibits temporal dependencies between observations, which can be challenging for traditional machine learning algorithms that assume independence. Covariates may change over time, necessitating their inclusion in the model. Moreover, these time-varying data often lack stationarity, potentially affecting model performance if not addressed. Lastly, 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.

Therefore, in this study, we proposed novel TSB estimators, based on mixed-effects random forests, incorporating specific physiological insights as additional mixed-effects terms. We evaluated the proposed models and we compared them to a standard random forest. It should be noted that in this study, we are not claiming to propose better alternative tools to TSB or TcB measurements. Instead, we aim to study the effectiveness of mixed-effects-based machine learning models in this context by comparing their performance in multiple aspects with those obtained from standard machine learning approaches.

4.2 Database

4.2.1 Study population

This study is based on the CARESS-Premi project, as described in Section 2.1.3, as an ancillary study proposed to estimate TSB levels using general clinical characteristics and longitudinal cardio-respiratory data. Here, a subset of the whole CARESS-Premi cohort admitted to the main clinical center (Rennes) was considered. The inclusion in this study covered the infants whose TSB levels had been measured at least once during follow-ups. TSB measurements taken after 34 weeks of postmenstrual age were dropped due to the clinical protocol requirements. Bilirubin levels higher than 400 µmol/L were excluded as well. This is based on the consideration that this kind of extremely high bilirubin level may be due to very complicated conditions associated with both immaturities with pathologies of the very premature infants and may not accurately reflect the typical clinical scenarios we aim to study, so it was therefore regarded as an exception. As a routine, during the stay in the NICU, infants were continuously monitored for their cardio-respiratory status, and the monitoring signals were de-identified and stored.

4.2.2 Data selection

In real medical and clinical settings, monitoring signals such as ECG are continuously monitored and collected, whereas bilirubin levels in infants are only measured sparsely due to the invasive nature of the blood sampling procedure. A data selection phase was thus implemented to align two data sources with very different sampling resolutions, thereby constructing a dataset suitable for subsequent analyses.

This data selection process involves a series of queries to the database and the synchronization of clinical events and associated clinical data with the monitoring signals. Figure 4.1 presents an illustration of the process. Additional information regarding the data structure in the CARESS-Premi database is introduced in Section 2.1.3.

During one selection, for instance, we first specified an instant (timestamp) at which TSB levels were measured from a patient based on the metadata file. Meanwhile, we queried all raw data files corresponding to this patient across all monitoring channels and sorted them chronologically. Subsequently, using the instant (TSB measurement timestamp) as an anchor, we extracted a predefined length of raw ECG recordings centered on this TSB measurement, specifically in our case a 4-hour segment (two hours before and two hours after the instant). Finally, the selected raw signals and metadata (such as postnatal age and postmenstrual age) related to a single bilirubin measurement for one patient were organized and stored in a newly created hierarchical data format file (HDF5).

This selection process was applied to all TSB measurements for all included infants. Yet, due to technical issues and other factors, some raw electrophysiological recordings were missing from the database. If data were unavailable around the anchor timestamp of TSB observations, this

would result in unsuccessful data selection and a certain of data loss. As a result, we obtained a set of 4-hour data segments, which are used for further processing and analysis.

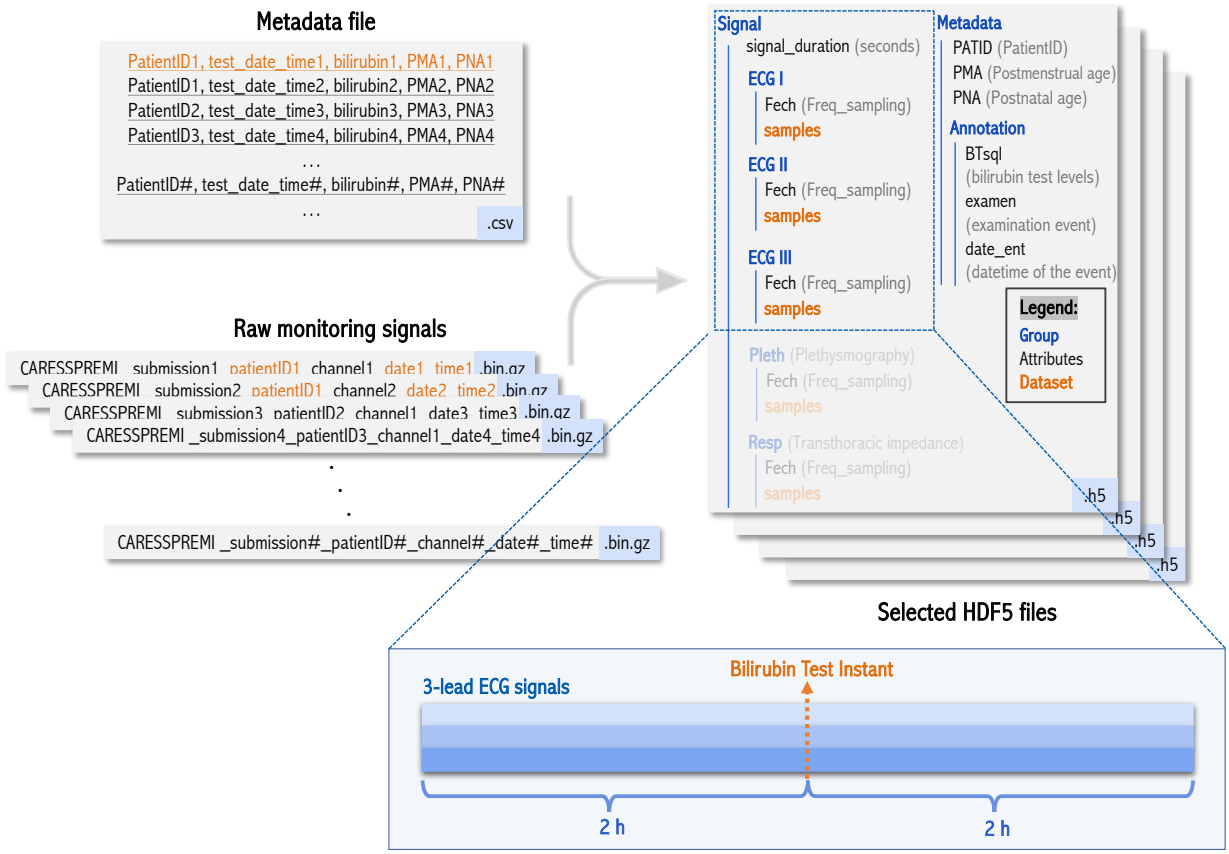

Figure 4.1: Data selection process including raw monitoring signal query, data synchronization and HDF5 file creation.

4.3 Signal Processing and Feature Engineering

The data processing generally follows our pipeline as introduced in Section 2.2, integrating ECG signal processing, time-series denoising, and feature engineering. A global representation of the data processing workflow is shown in Figure 4.2. The following sections describe each block of the workflow.

4.3.1 ECG signal processing and time-series denoising

Raw 3-channel ECG signals acquired based on each measurement, sampled at 500 Hz of frequency, were initially segmented into successive 15-minute windows with a 20% window overlap. We then chose the good lead of ECG from the multi-channel recordings using the algorithm described in Section 2.2.1. Each 15-minute ECG segment was pre-processed in order to detect cardiac

Figure 4.2: Data processing workflow for bilirubin estimation.

beats (QRS complex) using a robust multi-feature probabilistic real-time QRS detector adapted to the specific characteristics of the newborn electrophysiology, previously proposed and evaluated by our *SEPIA* team [13]. Once the QRS detection has been finalized, the successive cardiac cycle lengths, i.e., the time intervals elapsed between consecutive heartbeats, were calculated to build the so-called RR series.

Next, a multi-step approach using logic rules based on pathological and rhythmic corrections (details refer to Section 2.2.3) was used to automatically reject and correct artifacts and errors. Changes over time in the mean and variance of the corrected RR series were then estimated and signal stationarity was analyzed on each of the 15-minute segments (details refer to Section 2.2.4). By the fact that the events of interest are infrequent and present relatively slow dynamics, we assume that any segment of the 4-hour recordings we located could represent the physiological characteristics related to this event, i.e., TSB concentration. The signal quality and stationary were also taken into account. Therefore, only the most stationary segment in each 4-hour corrected RR series was selected for further analysis.

4.3.2 Feature extraction

Heart rate variability (HRV) features describing cardiovascular functions modulated by the Autonomic nervous system (ANS) were extracted from the most stationary RR series. Refer to HRV parameters in Section 2.2.5 for detailed descriptions.

The time domain analysis consists of the extraction of the mean (Mean), the median (Median), the standard deviation (Std), and the skewness (Skewness) as well as kurtosis (Kurtosis) of the RR series, which offers an initial indication of the global variability. And the square root of the mean squared differences of the successive RR (Rmssd), deceleration capacity (DC) and acceleration capacity (AC) based on the phase-rectified signal averaging method, the inter-decile range between 10^{th} and 90^{th} percentile (IDR), percentage deceleration of RR intervals (pDec), the standard deviation of RR corresponding to pDec (stdDec) and sample asymmetry of RR histogram (SampAsym), were also calculated.

The frequency domain analysis includes the integration of low-frequency (LF, 0.02-0.2 Hz) and high-frequency (HF, 0.2-2 Hz) ranges of the power spectrum obtained by autoregressive modeling

of the 4 Hz resampled RR series. Both features in normalized units (LFnu and HFnu, respectively) and the LF/HF ratio (LFHF) were also calculated. Very low-frequency variations (0-0.02 Hz) are not considered for the analysis of these short-duration segments.

Classical non-linear analysis such as SD1 and SD2 from the Poincaré plot, sample entropy (SampEn), and detrended fluctuation analysis (coefficients $\alpha 1$ and $\alpha 2$) were included in this study.

In total, we extracted 22 HRV parameters including 12 features in the time domain, 5 in the frequency domain, and another 5 non-linear features.

4.3.3 Outlier exclusion

We are coping with real-life data that contains various sources of noise, which can contaminate and distort the data. Additionally, after feature extraction, an outlier exclusion was hereby implemented to detect outliers in the feature space. We used a Chi-square test (α=0.05) on the Mahalanobis distance to identify the outliers in a multi-dimensional manner using (details refer to Section 2.3.4). These outliers are then excluded from the dataset.

4.4 Machine Learning Models for Clinical Longitudinal Data Analytics

Machine learning models are widely used in the field of healthcare. In the particular case of healthcare data acquired during a monitoring process, a number of specific challenges arise, which are associated with [14]: *i*) the time-dependent (longitudinal) aspect of these data, often exhibiting temporal dependencies between observations; *ii*) time-varying covariates that may change over time, making it essential to account for these changes in the model; *iii*) noises and missing values due to measurement errors or non-response rates, which can lead to biased estimates if not handled properly and *iv*) time-series non-stationarity. These aspects can be particularly challenging for traditional machine learning algorithms that assume independence.

In the following sections, we describe and propose a set of models that are progressively adapted to the particular characteristics of these monitoring data, taking into account potential mixed effects and including an explicit representation of the physiological knowledge related to the dynamics of TSB.

4.4.1 Baseline random forest (RF)

A standard RF regressor was employed for TSB estimation in the form of

$$y = f(\boldsymbol{X}) + \epsilon \tag{4.1}$$

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

However, like other machine learning approaches, the RF algorithm assumes that observations are independently sampled from a population. Analyzing longitudinal data without accounting for the inherent correlations between observations might result in biased inferences due to underestimated standard errors in linear models [15, 16].

4.4.2 Mixed-effects random forest (MERF)

The Mixed-Effects Random Forest (MERF), which originated from linear mixed-effects models, has been augmented to more effectively manage clustered and longitudinal data with repeated measurements within clusters/subjects by adding linear random effects into the models. Hajjem et al. [17] proposed the MERF in 2014, integrating a powerful and robust ensemble learning algorithm, formulated as

$$y = f(\boldsymbol{X}) + \boldsymbol{b}_i \boldsymbol{Z} + \boldsymbol{\epsilon} \tag{4.2}$$

where y is TSB measurements, $\boldsymbol{X}$ is a matrix of fixed-effects features, $f(\cdot)$ denotes a non-linear function, i.e., a standard RF regressor, $\boldsymbol{Z}$ is a matrix of random effects covariates specific to each subject capturing inter-subject variability, $\boldsymbol{b}_i \sim N(0, \sigma_b^2) = N(0, \boldsymbol{D})$ are random effect coefficients, should be estimated for each patient i and are assumed to be drawn from the same distribution where $\boldsymbol{D}$ is learned from the data, for each patient i (the cluster index), and $\boldsymbol{\epsilon} \sim N(0, \sigma_e^2) = N(0, \boldsymbol{R}_i)$ represents individual error. The covariance matrix of $\boldsymbol{b}_i$ is $\boldsymbol{D}$, the covariance matrix of $\boldsymbol{\epsilon}$ is $\boldsymbol{R}_i$ and σ_e^2 is the variance of $\boldsymbol{\epsilon}_i$ that is assumed to be white noise. An illustration of the model structure is shown in Figure 4.3.

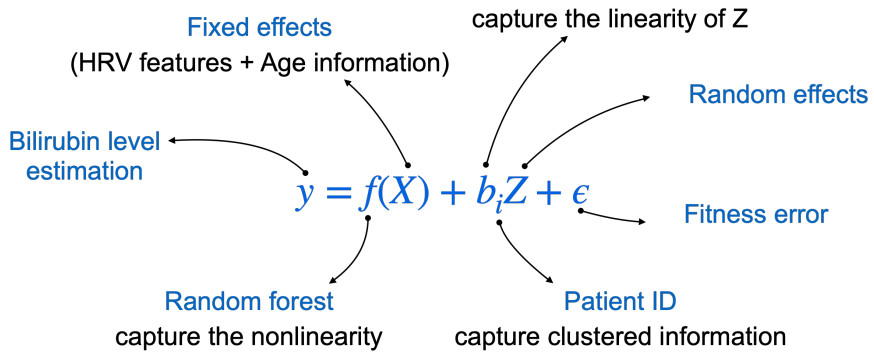

Figure 4.3: Mixed-effects random forest (MERF) model for bilirubin estimation.

The MERF is implemented under the framework of the Expectation-Maximization (EM) algorithm, which was originally used for linear mixed-effects models [18], and then adapted by Hajjem et al. [19] by replacing the non-linear function with a standard RF. The algorithm iterates until convergence, jointly learning random effect coefficients ($\boldsymbol{D}$), error priors (σ_e^2) and the RF ($f(\boldsymbol{X})$),

evaluating the Generalized log-likelihood (GLL) criterion at each iteration. Refer to Section 2.4.1 for more information on the algorithms.

4.4.3 Modified mixed-effects random forest (mMERF)

In the bilirubin characterization study as presented in Chapter 3, we observed that the natural evolution of TSB levels with PNA in the neonatal period of preterm infants follows an exponential decay pattern rather than any linear dynamics. Thus, inspired by the key idea of MERF that assumes the random effects are linear, we propose to make a modification to the MERF structure by replacing its linear term with a non-linear term while maintaining the optimization process untouched.

To capture this non-linear relationship explicitly, we proposed a Modified Mixed-Effects Random Forest (mMERF) model in form of:

$$y = f(\boldsymbol{X}) + b1_i(Z1_i) + b2_i(g(Z2_i)) + \epsilon \tag{4.3}$$

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 bilirubin dynamics:

$$g(\boldsymbol{Z2_i(t)}) = A \times exp(-B \cdot \boldsymbol{Z2_i(t)} + GA + Tc) + C \tag{4.4}$$

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 [20]. Figure 4.4 illustrates the model structure.

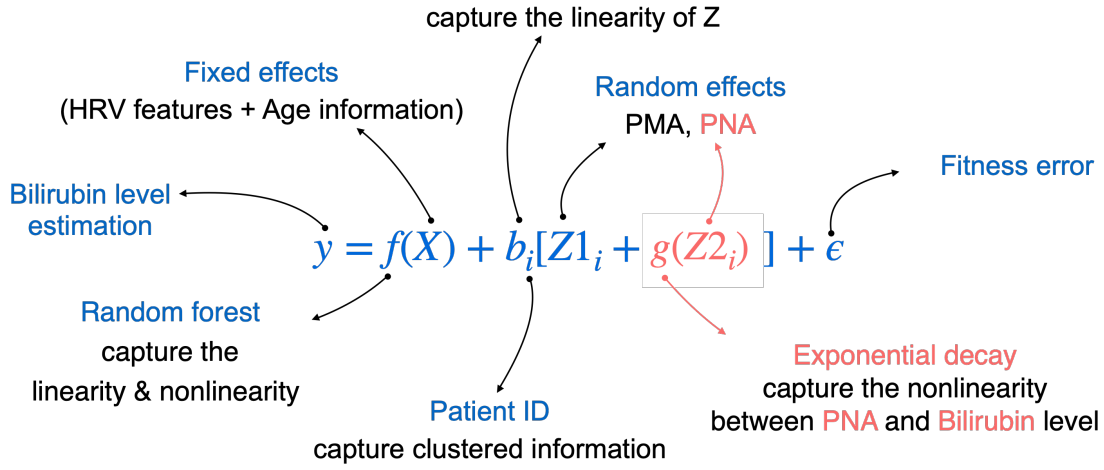

Figure 4.4: Modified Mixed-Effects Random Forest (mMERF) model for bilirubin estimation.

4.4.4 Model development

Models were learned 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.

In order to evaluate the added value of the proposed random-effects factors, we developed different mixed-effects-based variants:

- RF_{base}: a standard RF was learned and used as a baseline model (Equation 4.1).
- Based on the MERF structure, two models were proposed:
  - $MERF_0$: a pure MERF with no extra random effects but only assigning clusters (infants) for samples (Equation 4.2 with a sole $Z = \mathbf{1}$);
  - $MERF_2$: a MERF with two random effects to capture patient-level variability (Equation 4.2 with $Z1 =$ PMA and $Z2 =$ PNA).
- mMERF: a modified MERF was trained with a linear random effect of PMA and a non-linear random effect of PNA according to Equation 4.3.

Optimal hyper-parameters for baseline RF and RFs embedded in (m)MERFs were identified via Bayesian optimization (implemented by Python library of *Hyperopt-Sklearn* [21]) within predefined parameter spaces (details refer to Table 4.1). The maximum number of EM iterations to train (m)MERF models was set to 200. We utilized open-source Python packages of *scikit-learn* [22] and *MERF* [17] for the modeling process.

4.4.5 Model evaluation

Evaluating standard RF regressors involves using the trained model in inference mode to make predictions on the unseen test set, as a common way in machine learning models. For the fitted (m)MERF estimators, the way to make predictions depends on whether the patients/clusters to which the new samples in the test set have been seen during training, and work as follows:

- For new observations from patients seen during training ("known" patients), predictions comprise a population-averaged RF estimation with additional random-effects corrections;
- For samples from patients not seen during training ("unknown" patients), the estimator reverts to a standard RF learned by only fixed effects.

To compare the estimation performances, several metrics focusing on different aspects of the models were used. For comparing the correlation between the observed TSB levels and the estimated levels, we calculated the Pearson correlation and the Root-mean-square error (RMSE).

For evaluating the agreement between the observed TSB levels and the estimated levels, the Bland-Altman (B&A) analysis [23–25] was employed. The analysis calculates the mean difference (bias, differences between actual TSB and estimated TSB) and the Limits of agreement (LoA) to

determine the presence of estimation bias and to evaluate the consistency between the real TSB measurements and developed models' TSB estimations across the measurement range. The LoA is defined as the mean difference ± 1.96 times the standard deviation of the differences, representing the range within which 95% of the differences between the measurements are expected to lie, assuming a normal distribution of the differences. The B&A plots were created to visualize the differences against the means, providing clear and intuitive results for easier comparison.

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 [26].

4.5 Results

4.5.1 Data

Initially, 2,280 bilirubin samples from 374 preterm infants were eligible. After applying the inclusion and exclusion procedures and the following data selection process, a dataset consisting of raw monitoring signals and clinical information linking to 1,652 TSB measurements from 326 preterm infants was built for this study. The mean gestational age (GA) of this population was $28^{3/7}$ weeks (from $24^{2/7}$ to $31^{6/7}$ weeks). Their birth weights ranged from 520 to 1,955 grams with a mean of $1,132 \pm 316$ grams and the mean Z-scored birth weight [27] was -0.04 ± 0.78. In terms of TSB measurements, a total of 1,652 valid concentrations from the included infants were documented.

After feature engineering and outliers removal, 1,477 data samples from 319 infants eventually remained for model training and validation. The median number of TSB observations per patient was 4 (ranging from 1 to 22), and the mean TSB concentration was 118 ± 50 µmol/L.

A visualization of the data distribution across the analyzed population sorted by increasing PNA (in days) is illustrated in Figure 4.5, where the dots represent TSB measurements and the different color tones exhibit TSB levels (darker color for higher values). It is noticed from the number of scatters in each infant (each vertical set of scatters) that the more premature the newborn, the more frequent the blood tests to monitor bilirubin levels. As the gestational age increases, i.e., from extremely preterm to very preterm, the fewer the documented TSB measurements. Besides, from the color distribution in the scatter plot, we can see that most high TSB values are concentrated in the first days after birth (lower area of the plot), approximately the first 10 days of age. The overall trend of TSB concentrations decreases with increasing postnatal age, which is consistent with the perspective we presented in Chapter 3. Figure 4.6 displays the box plots of HRV parameters.

Regarding the machine learning data split, the training set comprised 1,181 samples covering 308 infants, and the test set included 296 samples, with 170 infants also presented in the training set and 11 infants being new to the models.

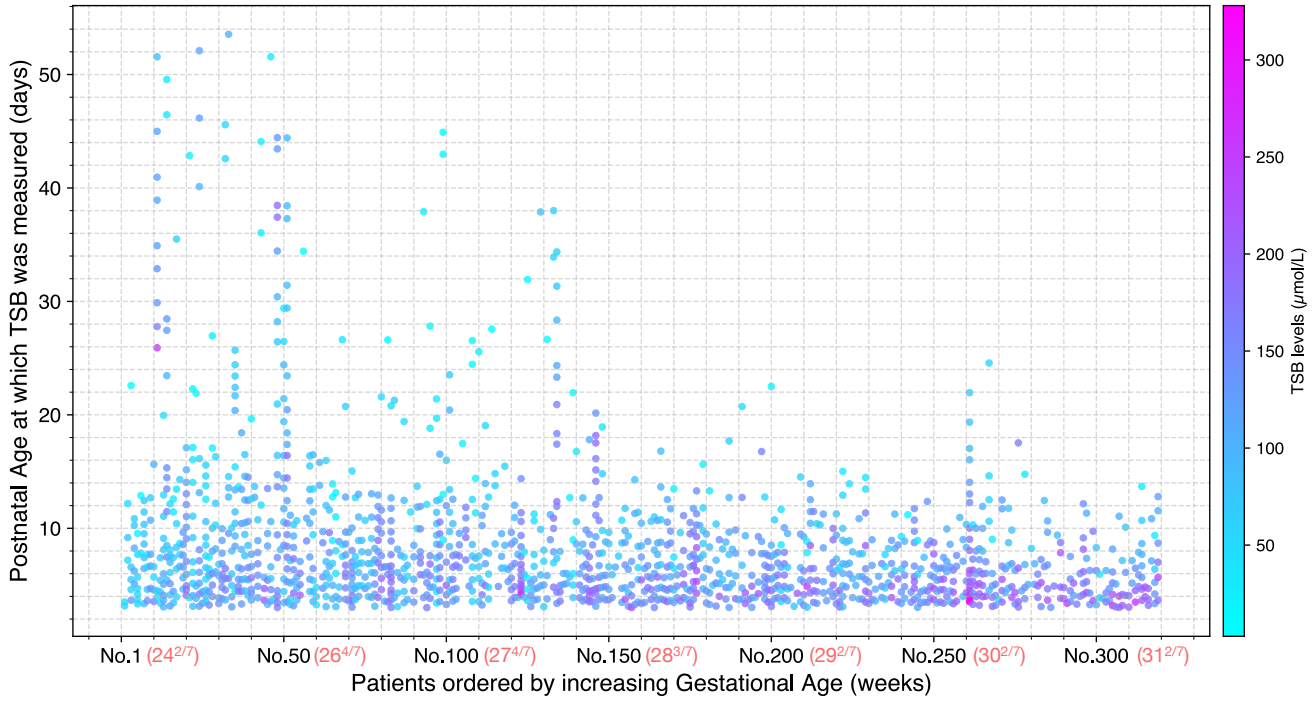

Figure 4.5: Distribution of data used for bilirubin estimation.

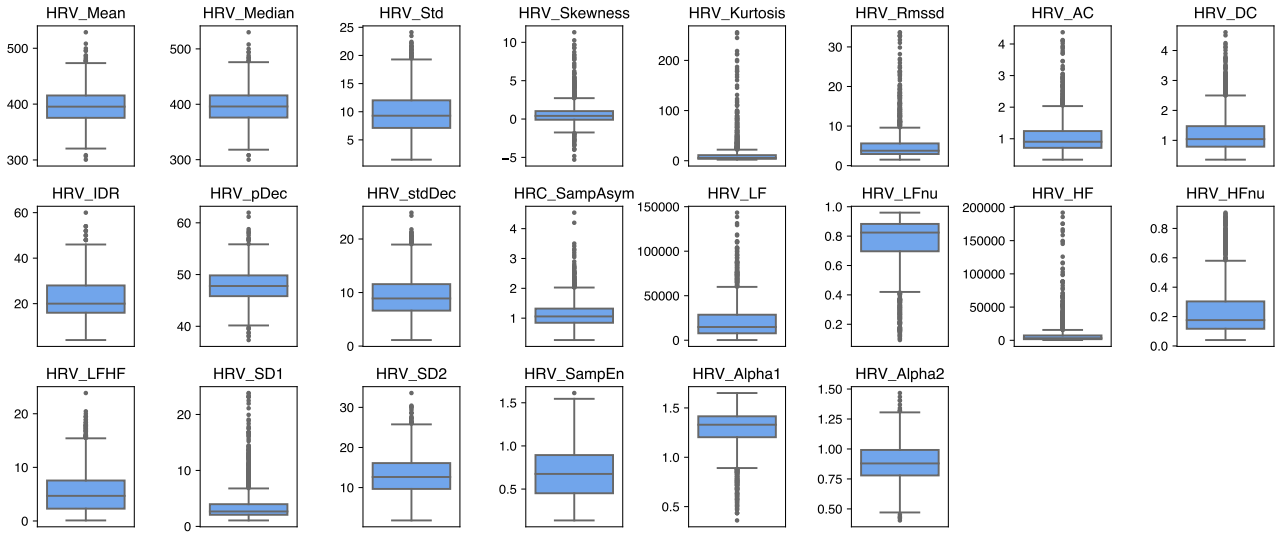

Figure 4.6: Box plots of HRV features used for bilirubin estimation.

4.5.2 Model comparison

Developed models

The first model was a standard random forest regressor (RF_{base}). The hyper-parameters of this model were optimized in the training set (details refer to Table 4.1), including 80 estimators (decision trees in the forest), *None* for the maximum features to consider when looking for the best split, 7 for the maximum depth of the trees, 2 samples for the minimum number required to be at a leaf node and 10 samples for the minimum number needed to split an internal node, and 0.05 for the threshold of decreased impurity of a split node, while the rest were set as default.

Table 4.1: Hyper-parameter space and the optimal choices for random forest estimator.

Hyper-parameter	Optimization Space	Optimal Choice
n_estimators	50, 60, 70, 80, 90, 100	80
max_features	4, 5, 6, 7, 8, 9, 10, 12, 14, 16, None	None
max_depth	3, 4, 5, 6, 7, 8, 9, 10, None	7
min_samples_leaf	1, 2, 3, 4, 5, 6	2
min_samples_split	2, 4, 6, 8, 10, 12	2
min_impurity_decrease	0, 0.01, 0.02, 0.05	0.05
criterion	Default	'squared_error'
bootstrap	Default	True

This set of hyper-parameters was also assigned to all the embedded random forest models as part of the (m)MERF variants for better comparison. As a result, we obtained four groups of models:

1. Baseline random forest model: RF_{base};
2. Basic mixed-effects random forest model and its embedded random forest term: $MERF_0$ and RF_{MERF_0};
3. Mixed-effects random forest model with 2 random effects and its embedded random forest term: $MERF_2$ and RF_{MERF_2};
4. Modified mixed-effects random forest model and its embedded random forest term: mMERF and RF_{mMERF}.

Training statistics of the (m)MERF models

Shown in Figure 4.7 are the training processes of the three mixed-effects-based models: $MERF_0$, $MERF_2$ and mMERF. Each subplot illustrates the Generalized log-likelihood (GLL), estimations of

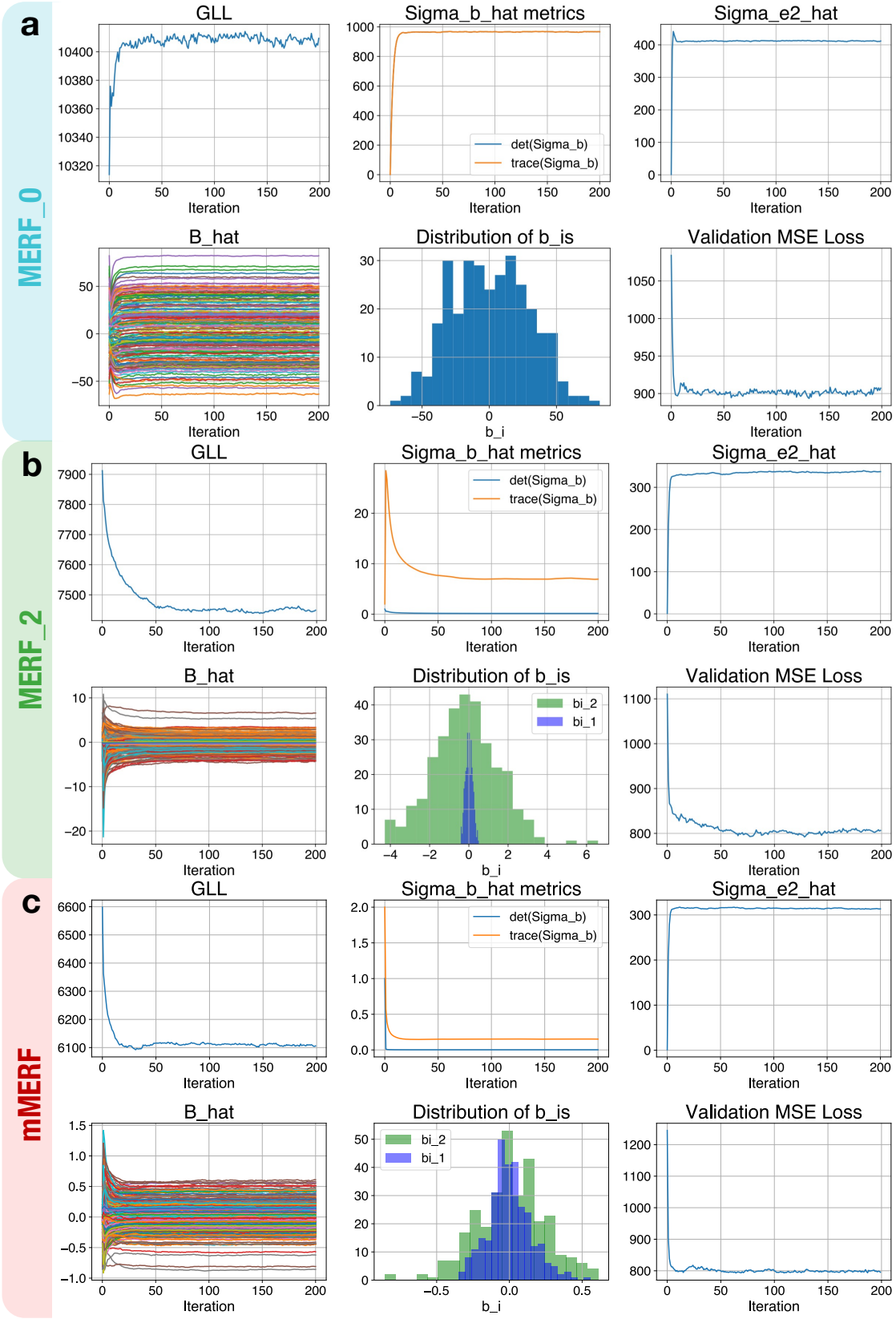

Figure 4.7: Evolution of the training statistics during 200 iterations for (**a**) the MERF_0 model, (**b**) the MERF_2 model and the (**c**) mMERF model.

σ_b and σ_e^2, the evolution and distribution of random coefficient $\boldsymbol{B}$ and the validation loss over 200 iterations.

Figure 4.7a depicts a relatively poor convergence of the MERF_0 model that the random effect was configured with a vector of ones, compared to subplots (**b**) and (**c**), which is indicated by the highest magnitude and jittered trace of GLL as well as the highest validation loss until the end of the iteration. And the "B_hat" as well as the distribution of "b_is" present the widest ranges. Curves in different colors of "B_hat" are different random coefficients for each cluster (infant), showing some initial fluctuations but generally stabilized around different values. Figure 4.7b shows the training process of the MERF_2 model, which includes two linear random effects: PMA and PNA in days. It had a slower but more robust convergence trajectory of GLL from around 50 iterations. And the random effects "B_hat" were estimated within a narrower value range of around [-5, 5]. Lastly, Figure 4.7c illustrates the mMERF model with a linear and a non-linear random effect (see Equation 4.3). During the first 20 iterations, there is a rapid decrease in GLL and evident changes in the estimated covariates of both σ_b and σ_e^2 and the validation loss. Afterward, the values stabilize with low-amplitude fluctuations around stable means, indicating the quick convergence of the algorithm. A more compact "B_hat" dispersion that is caused by greater magnitudes of two random effect factors illustrates similar global behaviors on the lines with the first two. Stabilization at different values reflects the models' adaptation to the individual differences among infants, accounting for variability in their data.

Model performances

Table 4.2 summarizes the overall models' performances and Figure 4.9 and Figure 4.10 display the B&A plots for the performances of the developed models in the test set and training set, respectively. In general, from RF_{base} to MERF_0, MERF_2 and finally the mMERF, the model performances improved as the expansion of the model structure and the increase of incorporated information.

As shown by the correlations and RMSE between the real and estimated TSB levels, the baseline model RF_{base} had a correlation of 0.53 and an RMSE of 42.41 µmol/L while the proposed three models achieved higher correlations (above 0.80) and lower errors (around 30 µmol/L). Figure 4.8 shows the distribution of the real and estimated bilirubin concentrations. As part of the (m)MERFs, $\text{RF}_{\text{MERF}_0}$, $\text{RF}_{\text{MERF}_2}$ and RF_{mMERF} had even poorer performance compared to the RF_{base} which is reasonable.

Regarding the results of Bland-Altman analysis, the RF_{base} had the best mean difference while the proposed three models slightly overestimated the bilirubin levels indicated by negative mean differences. We used a percentage of LoA relative to the LoA of RF_{base} model to compare different models, and a positive ΔLoA means a wider range of agreements suggesting poor consistency while a negative ΔLoA indicates improvement in consistency. Thus, as shown in Table 4.2, all the mixed-effects-based models, MERF_0, MERF_2 and mMERF, enhanced the agreements through

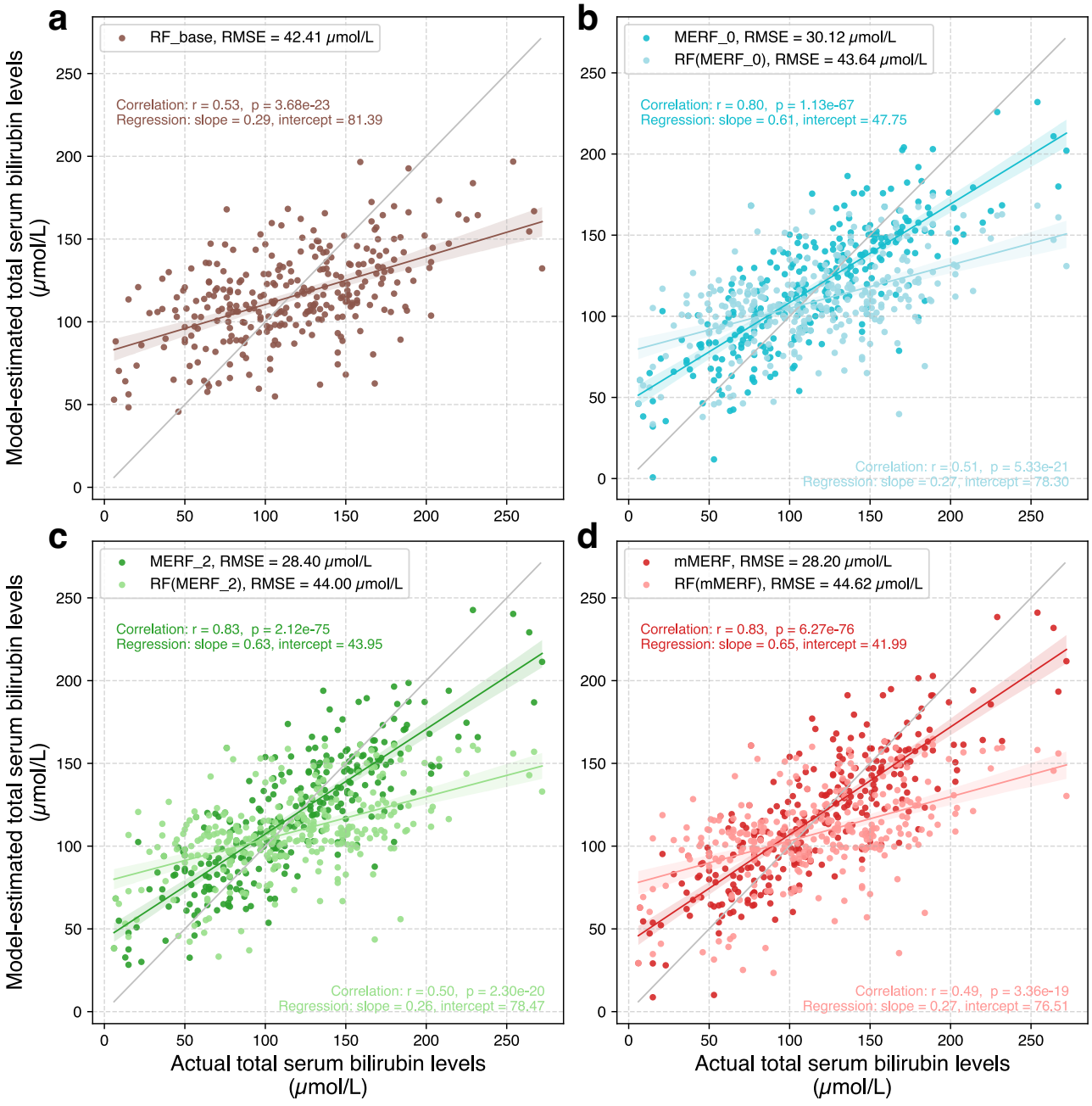

Figure 4.8: Actual TSB measurements against model-estimated TSB levels in the test set. (**a**) RF_{base} model, (**b**) $MERF_0$ model, (**c**) $MERF_2$ model and (**d**) mMERF model.

Table 4.2: Performance of the bilirubin estimation models in the test set.

Group	Model	Corr.†	RMSE (μmol/L)	mDiff (μmol/L)	LoA	ΔLoA	Corr. diff‡	Linear bias⋆
1	RF_{base}	0.53	42.41	0.82	83.10	/	0.61	0.76
2	RF_{MERF_0}	0.51	43.64	6.74	84.51	+1.70%	0.63	0.81
	$MERF_0$	0.80	30.12	−2.19	58.88	−29.15%	0.41	0.29
3	RF_{MERF_2}	0.50	44.00	7.65	84.92	+2.19%	0.64	0.83
	$MERF_2$	0.83	28.40	−1.53	55.58	−33.12%	0.42	0.27
4	RF_{mMERF}	0.49	44.62	8.49	85.87	+3.33%	0.59	0.77
	mMERF	0.83	28.20	−1.43	55.21	−33.56%	0.39	0.24

mDiff: Mean difference between the observed and model-estimated bilirubin levels.

LoA: 95% limits of agreements in the Bland-Altman plots.

Δ*LoA:* Changes in LoA relative to the baseline model RF_{base}.

†Pearson correlation between the observed and model-estimated bilirubin levels.

‡Pearson correlation between the mean and differences of observed and model-estimated bilirubin levels of samples from "known" patients.

⋆The slope in linear regression of samples from "known" patients based on the Bland-Altman plots.

narrowing LoA by 29.15%, 33.12% and 33.56%, respectively in the test data (41.36%, 49.83% and 52.90% in the training data). RF_{MERF_0}, RF_{MERF_2} and RF_{mMERF} largely underestimated the bilirubin levels (higher mean differences) and produced poorer agreements (wider LoA and positive ΔLoA).

Moreover, we calculated the correlation between the means and differences of observed and model-estimated TSB levels and, if correlated (statistically significant), performed a linear regression of these two variables to quantify the proportional bias of the estimators. The results show that, compared to RF_{base} that obtained a correlation of 0.61 and slope of the linear bias of 0.76, the proposed $MERF_0$, $MERF_2$ and mMERF models reduced the correlation to around 0.40 and greatly attenuated the undesirable bias by decreasing the linear regression slopes for the samples from "known" babies in the test set from 0.76 to 0.29, 0.27 and 0.24, respectively (from 0.53 to 0.18, 0.15 and 0.13 in the training set, respectively). Concerning RF_{MERF_0}, RF_{MERF_2} and RF_{mMERF} models, they presented similar or marginally higher correlation and proportional bias to the RF_{base} model, suggesting even poorer performance.

In summary, as shown by both the values in Table 4.2 and the visual representations in Figure 4.9, the standard model RF_{base} exhibited wider dispersion and noticeable proportional bias in both the training and test data; in contrast, the three proposed mixed-effects models—$MERF_0$, $MERF_2$ and mMERF—demonstrated significant improvements in estimation accuracy. This en-

hancement is rather evident in the Bland-Altman plots displayed as percentages, provided in Appendix Figure A.9 for the training data and Figure A.10 for the test data. By plotting the differences as percentages, it becomes clear that the variability of the bias decreases as TSB measurements increase, and that the degree of dispersion in the differences is greatly diminished in the mMERF model compared to the RF_{base} model. This suggests that the proposed models greatly improved TSB estimations, particularly for higher concentrations.

4.6 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. Classical machine learning has been widely used in developing prediction models with longitudinal data in healthcare, assisting in the prevention, diagnosis, and prognosis anticipation across various conditions and scenarios. Among these methods, Random Forest (RF) is one of the state-of-the-art and representative non-parametric ML approaches adept at handling both classification and regression tasks. It is capable of working with predictors of various scales or distributions and is suited for applications in high-dimensional settings, which exactly cater to the features of biomedical data.

Yet, as with many ML algorithms, RF analyzes data without considering the dependency among observations in longitudinal data, which could lead to great biases [16]. In contrast, methods designed to align with the inherent data structure and appropriately address the correlations arising from repeated measurements have been shown to deliver superior prediction performance [28]. This is where Mixed-Effects Random Forest (MERF) comes into play. MERF combines the strengths of random forests with mixed-effects models, thereby incorporating the ability to handle longitudinal data's hierarchical structure so that it can effectively manage the correlation within repeated measures and gain more reliable predictions.

In this study on bilirubin estimation using monitoring signals, though the task is fundamentally a regression problem, we observed how the samples distributed as shown in Figure 4.5 and this led us to build models that can learn across the population, but, at the same time, can account for the idiosyncrasies of each infant. We hypothesized that incorporating mixed effects that capture inter-patient characteristics into machine learning could significantly enhance model performance.

We developed MERF models with different combinations of random effects and compared their contributions added in TSB estimation when taking a standard RF model as a baseline. The results show incremental improvements from RF_{base} to $MERF_0$ and $MERF_2$, and subsequently to the mMERF model regarding higher correlations, greater agreements and less proportional bias, suggesting their advantages in analyzing longitudinal clinical data with patient-level repeated measurements.

The training trajectories of "B_hat" and the distributions of "b_{is}" in Figure 4.7 reveal interest-

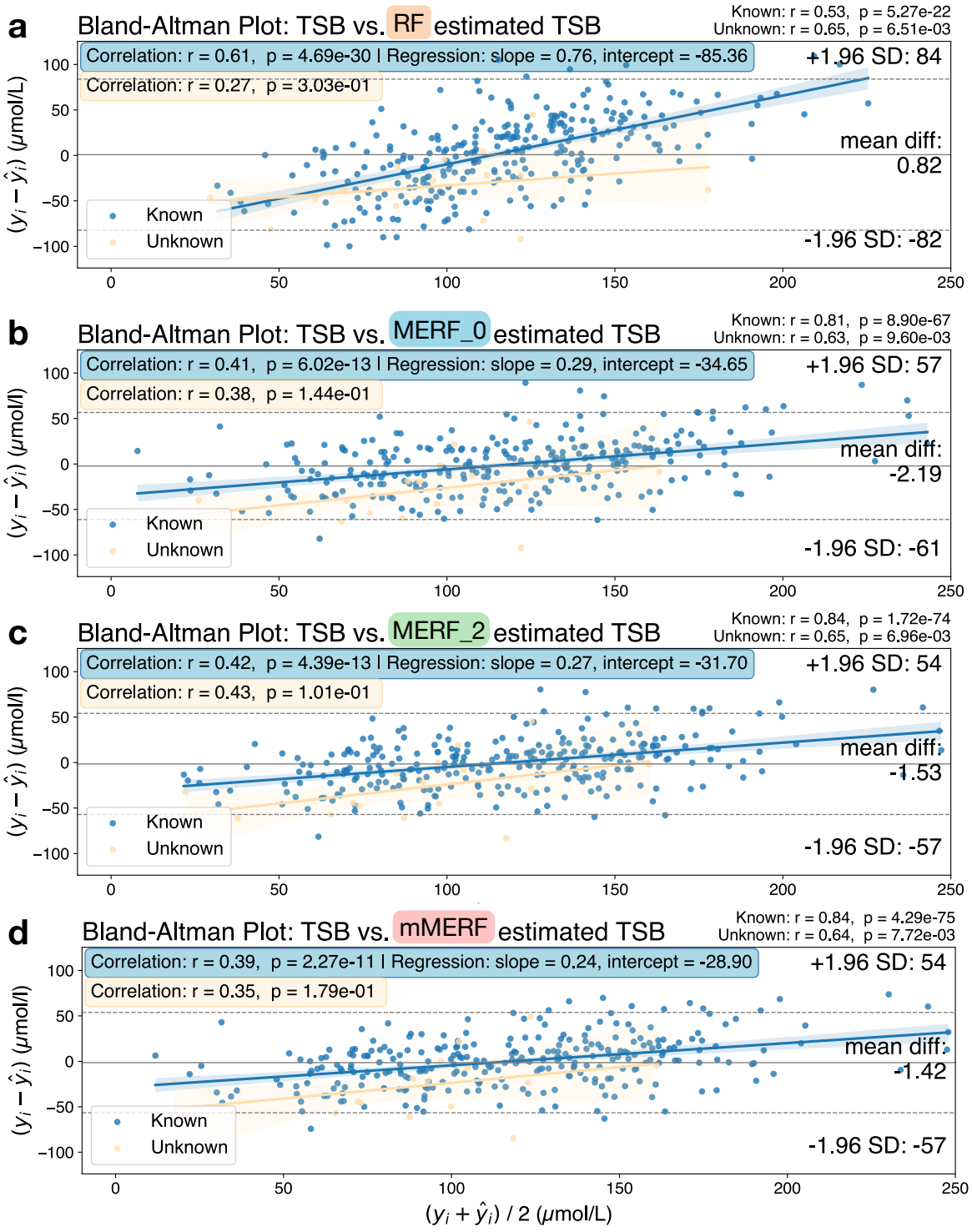

Figure 4.9: Bland-Altman plots of the mean against the differences between observed (y_i) and estimated ($\hat{y}_i$) TSB levels for samples of "known" patients (blue) and samples of "unknown" patients (beige) in the test set. (**a**) RF_{base} model. (**b**) $MERF_0$ model. (**c**) $MERF_2$ model. (**d**) mMERF model. The r and p on the upper right of each subplot indicate the Pearson correlations and associated p-values between the real TSB levels and the model-estimated TSB values.

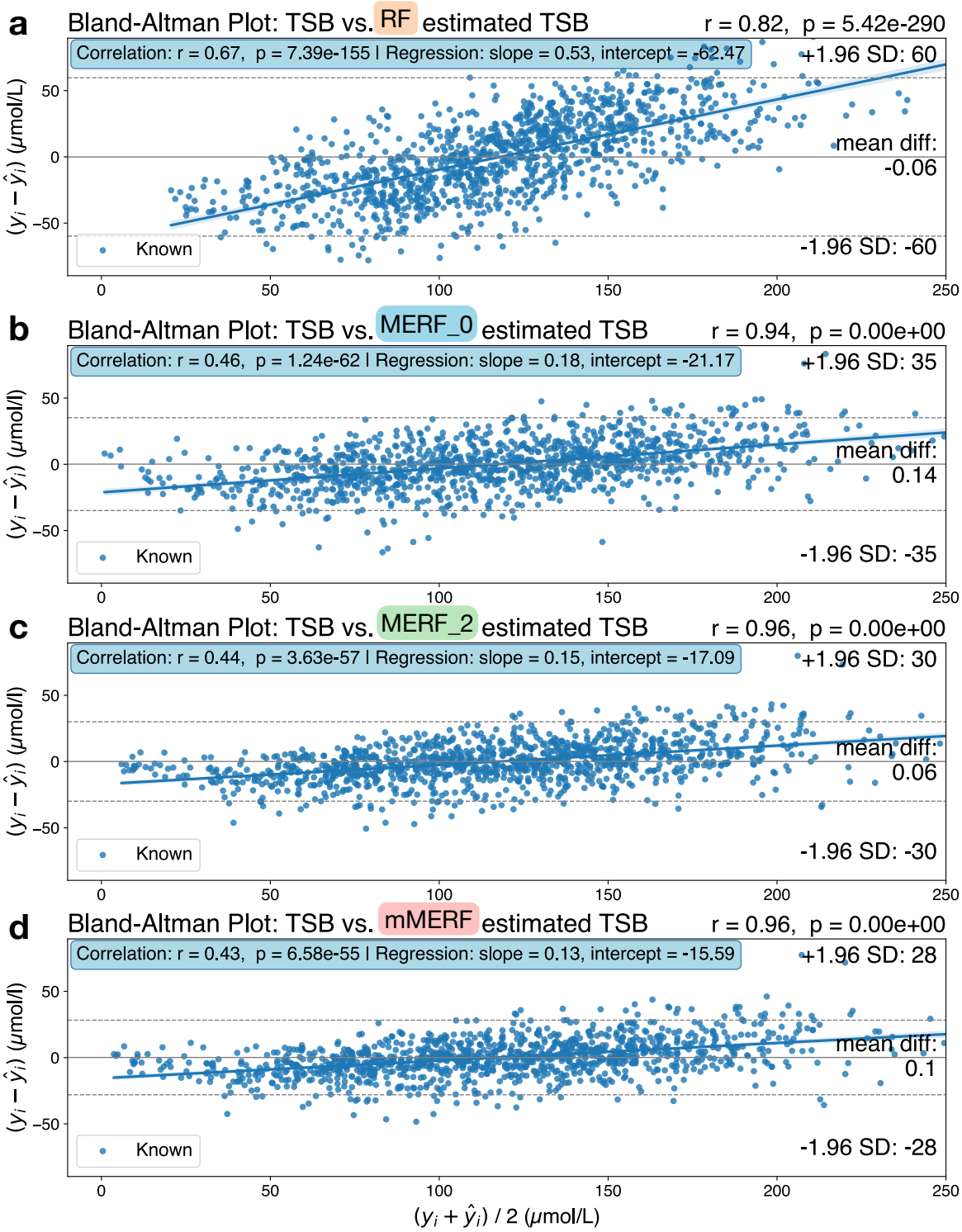

Figure 4.10: Bland-Altman plots of the mean against the differences between observed (y_i) and estimated ($\hat{y}_i$) TSB levels in the training set. (**a**) RF_{base} model. (**b**) $MERF_0$ model. (**c**) $MERF_2$ model. (**d**) mMERF model. The r and p on the upper right of each subplot indicate the Pearson correlations and associated p-values between the real TSB levels and the model-estimated TSB values.

ing information about the random effects on patient-specific variability. The $MERF_0$ model was configured with white noise random effect coefficients by only telling the model which patients the training samples belong to, which significantly increased estimation accuracy. Building on this, we further included two important clinical features known to be related to bilirubin levels into the $MERF_2$ model as two random effects: PMA and PNA, in order to integrate personalized prior knowledge about their impact on bilirubin dynamics. Due to the different scales of PMA and PNA values, the distributions of their random effect coefficients show different ranges (histograms in Figure 4.7b), but overall they are much more concentrated than the distribution of "b_{is}" for the $MERF_0$ model (Figure 4.7a). This can be explained by the fact that when more key information is explicitly integrated into the random effect terms, the corresponding random effect covariates can only capture less hidden information, which is manifested as smaller variances of "b_{is}" and narrower distributions. To further refine the model, we then proposed the mMERF model that replaces the second random effect of PNA with an exponential decay function of PNA, which breaks the limitation of MERF models that only consider linear random effects. Through this modification, an even quicker and robust convergence with lower validation loss can be seen in Figure 4.7c. This adjustment acknowledges the complex, nonlinear nature of bilirubin dynamics over time despite that a median model of 72 infants' bilirubin dynamics (Equation 3.6), rather than patient-specific models, was used. In this case, although the estimation performance is not evidently improved, the distribution of the random effects covariates is more compact. We assume that when more powerful, complicated and indicative insights are incorporated as random effect terms in MERF models, they will contribute more and possibly gradually dominate the output. In fact, the gradually decreasing performance of the random forests embedded in the (m)MERFs listed in Table 4.2, denoted as RF_{MERF_0}, RF_{MERF_2} and RF_{MERF}, offers us a glimpse into the subtle dynamic balance between the contributions of RF and random effect terms in a MERF model.

From the $MERF_2$ to the mMERF model, the big difference lies in the second random effect. In mMERF, instead of simply using the PNA corresponding to each TSB concentration measured, we included a mathematical deterministic model of the bilirubin evolution with PNA, as fully elaborated in Chapter 3. We expect this nonlinear effect in mMERF may reflect the complex nature of bilirubin changes, where the influence of factors such as postnatal age was proved not uniformly linear [29]. The results turned out to be somewhat disappointing, with the incorporation of such a function having only a negligible added value on the estimation of bilirubin levels, which was no different from the MERF models. We consider two potential aspects that might have affected the outcome. First, the exponential decay model used to represent bilirubin dynamics describes the natural course of bilirubin levels. As presented in Section 3.2.1, we excluded bilirubin measurements taken under conditions that could affect the natural evolution of bilirubin in neonates. However, in this study, we did not apply such strict exclusion criteria and instead evaluated all bilirubin levels using a set of non-invasive HRV and clinical features. This inconsistency might have resulted in the minimal improvement in bilirubin estimation seen with the addition of the nonlinear function. Second, while we have emphasized the importance of inter-subject variability

in longitudinal data, we did not incorporate patient-specific bilirubin evolution models, such as those modeled in Chapter 3. Instead, we used a general function composed of the median values of parameters from individualized evolution models. Due to the limited sample sizes in this study, it was challenging to have sufficient samples for each infant in the training set to properly fit a patient-specific exponential decay function. Consequently, we did not go further into this step. However, it is conceivable that once with adequate data, such an idea could promisingly enhance the performance of bilirubin level estimation non-invasively.

It should be noted that in this study, we do not claim to propose better alternative tools to TSB or TcB measurements. Instead, we aim to study the effectiveness of mixed-effects-based machine learning models in this context by comparing their performance in multiple aspects with those obtained from standard machine learning approaches. Even though, by observing the results of all the developed models (either RF or MERF), it can be noticed that, although better performances were obtained by incorporating different informative components into the models, there are still non-negligible estimation errors and proportional biases between model-estimated and the measured TSB levels. This drives us to wonder whether there is a strong enough association between changes in HRV and TSB levels of TSB to support the use of HRV parameters as main predictors for non-invasive estimation of bilirubin. Several studies have shown the relationship between hyperbilirubinemia and cardio-respiratory activities. For instance, the non-linear indices of HRV were proved associated with a decreased sympathetic activity and/or increased parasympathetic activity in full-term jaundiced newborns [9]. Also, as the first attempt to explore the potential link based on analyzing 24-hour Holter recordings, Ozdemir et al. have found significant differences in several HRV features in a population of full-term neonates with severe unconjugated hyperbilirubinemia, including higher RMSSD and HF and lower LF/HF ratio [11]. Research in preterm lambs suggested the correlation between moderate and sustained hyperbilirubinemia and altered cardio-respiratory function indicated by both respiratory rate variability and heart rate variability [10, 12]. However, whether a consistent association is also valid in preterm human infants and whether HRV analysis can effectively reflect changes in both normal and abnormal bilirubin concentrations of bilirubin still needs further investigation.

Some limitations of this study should be mentioned, as well as the prospects for future research directions that arise from them. Firstly, while the (m)MERF models showed greater consistency and less bias in bilirubin estimation for "known" patients, there is no significant difference in performance over the baseline RF_{base} model when encountering "unknown" patients in the test set (see orange scatters and lines in Figure 4.9). This is due to the inherent nature of mixed-effects models, which require patient-specific historical data to initialize the random coefficients. As mentioned above, limited by the small sample size, the "new" patients split in the test set but not seen in the training set failed to have corresponding random effect corrections when generating an estimation. Further improvements could involve more systematic and extensive data collection to build larger datasets. This would also facilitate more nuanced and personalized modeling to better

accommodate the heterogeneity present in longitudinal data observed in bilirubin monitoring.

Another point is about the exploration of random effects. As mentioned in the former discussion, the random effects in the proposed models might not be the optimal choices. Incorporating additional physiological insights, including patient-specific random effects, such as replacing $g(\cdot)$ with $g_i(\cdot)$ (Equation 4.4), may benefit to better encompass other bilirubin-related individual characteristics. This could improve the model's ability to accurately capture the individualized patterns of bilirubin dynamics. Apart from this, we would also like to mention the non-linear random effect but in a more methodological aspect. We conducted our studies based on an open-source Python package of *MERF* developed by Ahlem Hajjem, Francois Bellavance, and Denis Larocque [17]. One key assumption of the MERF model is that the random effect is *linear* and thus the *MERF* package was implemented in a way that the underlying EM algorithm alternatively optimizes the non-linear fixed effects model (an RF) and the linear random effect covariates ($\boldsymbol{b}_i$). In this study, we adopted a "lazy" way to realize our non-linear random effect by applying the exponential decay function before starting the optimization process. But a more elegant approach that could adapt to individualized non-linear random effects is to modify the EM algorithm, rewriting the underlying optimization steps and specific implementations, which is quite ambitious yet feasible.

A third limitation concerns the difficulty of converting such proposed models into On-the-edge (OTE) applications. The proposed models do not operate in real-time; they require historical data for each patient to learn and adjust the random coefficients effectively. This dependency on pre-existing data limits their immediate applicability to new patients who do not have sufficient historical records. Since, unfortunately, this is the inherent nature of a MERF structure, rather than optimizing the MERF models, further studies might involve exploring other advanced algorithms capable of quick adaptation or hybrid approaches that combine pre-trained models with real-time learning capabilities. Additionally, as an overall challenge, the integration of such models into existing neonatal care protocols and systems may require significant adaptations such as robust infrastructure to support real-time data processing and model updating. We will discuss this in Chapter 6.

4.7 Conclusion

In this chapter, we explored non-invasive approaches for estimating total bilirubin concentrations in preterm infants with/without hyperbilirubinemia born at $24^{2/7}$ to $31^{6/7}$ gestational weeks. A particular effort has been made to evolve a classical ML regressor into new structures adding specific terms to account for the longitudinal and problem-specific nature of the observed variables. This is conducted by developing original mixed-effects-based machine learning models through a set of heart rate variability and clinical features, especially incorporating personalized and knowledge-based random effects. The proposed mixed effects-based (m)MERF models (MERF_2 and mMERF) greatly improved the agreements and reduced the proportional bias thanks

to the explicit integration of meaningful physiological knowledge. This underscores the importance of considering data dependency and individual disparities in prediction modeling toward longitudinal clinical data in similar biomedical scenarios.

Although these models require patient-specific historical data for initialization and the model performance is still far from being compatible with an actual clinical application, with larger datasets and broader validation, this study proposes the continued optimization and integration of such advanced models into NICU monitoring systems, which may enhance real-time TSB estimation and optimize interventions for hyperbilirubinemia. Prospectively, the combination of a robust and adaptive platform that supports the deployment of such models into clinical decision support systems will hold promise to offer an optimistic step towards personalized and less invasive neonatal care.

BIBLIOGRAPHY

[1] L. M. Gartner, R. N. Snyder, R. S. Chabon, and J. Bernstein, "Kernicterus: high incidence in premature infants with low serum bilirubin concentrations," Pediatrics, vol. 45, no. 6, pp. 906–917, 1970.

[2] R. Stocker, Y. Yamamoto, A. F. McDonagh, A. N. Glazer, and B. N. Ames, "Bilirubin is an antioxidant of possible physiological importance," Science, vol. 235, no. 4792, pp. 1043–1046, 1987.

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

[4] S. M. Shapiro, "Definition of the clinical spectrum of kernicterus and bilirubin-induced neurologic dysfunction (BIND)," Journal of perinatology, vol. 25, no. 1, pp. 54–59, 2005.

[5] L. Johnson and V. K. Bhutani, "The clinical syndrome of bilirubin-induced neurologic dysfunction," Seminars in Perinatology, vol. 35, no. 3, pp. 101–113, 2011.

[6] J. F. Watchko, "Bilirubin-induced neurotoxicity in the preterm neonate," Clinics in Perinatology, vol. 43, no. 2, pp. 297–311, 2016.

[7] S. N. El-Beshbishi, K. E. Shattuck, A. A. Mohammad, and J. R. Petersen, "Hyperbilirubinemia and transcutaneous bilirubinometry," Clinical chemistry, vol. 55, no. 7, pp. 1280–1287, 2009.

[8] L. Ten Kate, T. van Oorschot, J. Woolderink, S. Teklenburg-Roord, and J. Bekhof, "Transcutaneous bilirubin accuracy before, during, and after phototherapy: A meta-analysis," Pediatrics, vol. 152, no. 6, p. e2023062335, 2023.

[9] Z. Uhrikova, M. Zibolen, K. Javorka, L. Chladekova, and M. Javorka, "Hyperbilirubinemia and phototherapy in newborns: Effects on cardiac autonomic control," Early Human Development, vol. 91, no. 6, pp. 351–356, 2015.

[10] M.-L. Specq, M. Bourgoin-Heck, N. Samson, F. Corbin, C. Gestreau, M. Richer, H. Kadhim, and J.-P. Praud, "Moderate hyperbilirubinemia alters neonatal cardiorespiratory control and induces inflammation in the nucleus tractus solitarius," Frontiers in Physiology, vol. 7, p. 437, 2016.

[11] R. Özdemir, Ö. Olukman, C. Karadeniz, K. Çelik, N. Katipoğlu, M. Muhtar Yılmazer, Ş. Çalkavur, T. Meşe, and S. Arslanoğlu, "Effect of unconjugated hyperbilirubinemia on neonatal autonomic functions: evaluation by heart rate variability," The Journal of Maternal-Fetal & Neonatal Medicine, vol. 31, no. 20, pp. 2763–2769, 2018.

[12] S. Al-Omar, V. Le Rolle, N. Samson, M.-L. Specq, M. Bourgoin-Heck, N. Costet, G. Carrault, and J.-P. Praud, "Influence of moderate hyperbilirubinemia on cardiorespiratory control in preterm lambs," Frontiers in Physiology, vol. 10, p. 468, 2019.

[13] M. Doyen, D. Ge, A. Beuchée, G. Carrault, and A. I. Hernández, "Robust, real-time generic detector based on a multi-feature probabilistic method," PLoS ONE, vol. 14, no. 10, p. e0223785, 2019.

[14] A. Cascarano, J. Mur-Petit, J. Hernandez-Gonzalez, M. Camacho, N. de Toro Eadie, P. Gkontra, M. Chadeau-Hyam, J. Vitria, and K. Lekadir, "Machine and deep learning for longitudinal biomedical data: a review of methods and applications," Artificial Intelligence Review, vol. 56, no. Suppl 2, pp. 1711–1771, 2023.

[15] S. W. Raudenbush and A. S. Bryk, Hierarchical linear models: Applications and data analysis methods. sage, 2002, vol. 1.

[16] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," Briefings in Bioinformatics, vol. 24, no. 2, p. bbad002, 2023.

[17] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed-effects random forest for clustered data," Journal of Statistical Computation and Simulation, vol. 84, no. 6, pp. 1313–1328, 2014.

[18] H. Wu and J.-T. Zhang, Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches. John Wiley & Sons, 2006.

[19] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed effects regression trees for clustered data," Statistics & probability letters, vol. 81, no. 4, pp. 451–459, 2011.

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

[21] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-Sklearn: Automatic hyperparameter configuration for scikit-learn." in Scipy, 2014, pp. 32–37.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[23] D. G. Altman and J. M. Bland, "Measurement in medicine: The analysis of method comparison studies," Journal of the Royal Statistical Society Series D: The Statistician, vol. 32, no. 3, pp. 307–317, 1983.

[24] J. Martin Bland and DouglasG. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," The Lancet, vol. 327, no. 8476, pp. 307–310, 1986.

[25] D. Giavarina, "Understanding Bland Altman analysis," Biochemia Medica, vol. 25, no. 2, pp. 141–151, 2015.

[26] K. M. Ho, "Using linear regression to assess dose-dependent bias on a Bland-Altman plot," Journal of Emergency and Critical Care Medicine, vol. 2, no. 8, 2018.

[27] T. R. Fenton and J. H. Kim, "A systematic review and meta-analysis to revise the Fenton growth chart for preterm infants," BMC Pediatrics, vol. 13, no. 1, p. 59, 2013.

[28] A. A. Mangino and W. H. Finch, "Prediction with mixed effects models: a monte carlo simulation study," Educational and Psychological Measurement, vol. 81, no. 6, pp. 1118–1142, 2021.

[29] N. Kawade and S. Onishi, "The prenatal and postnatal development of UDP-glucuronyltransferase activity towards bilirubin and the effect of premature birth on this activity in the human liver," The Biochemical Journal, vol. 196, no. 1, pp. 257–260, 1981.

CHAPTER

5 Early Detection of Neonatal Late-onset Sepsis in Preterm Infants Using Heart Rate Variability from Real-life Monitoring Data

This chapter considers another major challenge in the NICU: neonatal Late-onset sepsis (LOS) management. In line with the goals of the CARESS-Premi project, this work focuses on developing new computer-assisted diagnostic tools to support caregivers in the early detection and timely intervention of neonatal sepsis in preterm infants. This chapter explores the use of Heart rate variability (HRV) data from real-life monitoring data to enhance the precision and speed of sepsis detection, employing non-invasive, continuous bedside monitoring. This chapter begins with literature reviews covering previous studies on neonatal LOS, the use of HRV analysis in sepsis detection, and the progression of infections coupled with clinical intervention. Following this, the study population and expert classification of clinical events are introduced, which form the basis for annotating the dataset. Then, we detail the dataset construction process, from signal processing, and feature extraction to sample labeling. Subsequently, machine learning model development and evaluation methodologies for LOS detection are outlined. Finally, results are presented, along with discussions on several aspects of such a challenging topic—neonatal sepsis detection and its clinical application.

5.1 Introduction

5.1.1 Neonatal late-onset sepsis

As presented in the previous chapters (Section 1.3.2), neonatal Late-onset sepsis (LOS), defined as sepsis occurring after 72 hours of life, remains a critical global public health concern and is one of the leading causes of neonatal morbidity and mortality [1]. Preterm infants are particularly vulnerable to LOS due to their underdeveloped immune systems, longer hospital stays, and exposure to invasive procedures necessary for routine care. This heightened susceptibility often results in prolonged hospitalizations, higher healthcare costs, and an increased risk of long-term neurodevelopmental impairments.

Prompt diagnosis and the initiation of appropriate antibiotic therapy can significantly reduce mortality and morbidity [2, 3]. Given the rapid deterioration of infection in the preterm population, clinicians are often compelled to empirically administer antibiotics to those with risk factors or signs of suspected sepsis. However, this approach, sometimes aggressive and indiscriminate, carries the risk of overuse of antibiotics, which can lead to the development of antibiotic resistance, adverse drug-related consequences, and further healthcare costs [4–7].

Therefore, the early and accurate recognition of LOS is crucial to guide appropriate antibiotic treatments and improve clinical outcomes in this vulnerable population. However, the LOS early recognition is particularly challenging due to the non-specific and subtle clinical presentations in preterm infants [1]. Despite being the gold standard in traditional diagnostic tools, blood cultures are invasive, time-consuming, and can show variability in predictive accuracy, particularly in the early stages of infection [8].

As a result, there is a pressing need for non-invasive, timely, and reliable diagnostic methods to facilitate the detection of LOS as early as possible. One such promising approach is Heart rate variability (HRV) analysis, which has gained attention for its potential role in early sepsis detection. The following section provides a review of current literature on HRV analysis, highlighting its applications and effectiveness in identifying sepsis in neonates.

5.1.2 HRV analysis in sepsis early detection

Heart rate (HR) is a widely used physiological parameter in clinical practice and pediatric early warning systems [9] as an early marker of sepsis-related deterioration. However, there is a lack of consensus among warning scores on what constitutes an abnormal heart rate in pediatric patients [10]. This inconsistency has prompted the exploration of novel and more comprehensive methods, such as Heart rate variability (HRV) analysis.

HRV, a measure of the variation in time intervals between heartbeats (RR intervals), reflects the autonomic nervous system's regulation of cardiovascular function. In sepsis, changes in HRV are linked to autonomic dysfunction caused by systemic inflammation. Studies across neonatal, pediatric, and adult populations have shown promising results in using HRV analysis for sepsis management, including early detection and outcome prediction. Particularly, benefiting from continuous non-invasive cardiac monitoring is often a standard procedure in the NICU, studies have explored the use of HRV analysis as a diagnostic and prognostic tool for neonatal sepsis.

Pioneering work in this domain was conducted by Griffin, Moorman and their colleagues, who introduced a novel and proprietary measure known as Heart rate characteristics (HRC), to evaluate HRV in infants at risk of sepsis. The main subjects were neonates admitted in the NICU at the University of Virginia. Their research indicated that abnormal HRC, characterized by reduced variability, transient decelerations and increased non-stationarity, 12 to 24 hours preceded the on-

set of neonatal sepsis [11–20]. A key finding from Griffin and colleagues' research is that HRC measurements, which involve continuous monitoring and analysis, provide independent and complementary information to conventional demographics (gestational age, birth weight, and days of age) [13], invasive and time-consuming laboratory tests and vital sign assessments [17] regarding the prediction of neonatal sepsis and death.

Griffin et al. proposed an HRC index that incorporates various statistics and measures: HR changes (standard deviation), asymmetry of histograms of RR intervals (sample asymmetry [21]), and repeating patterns of RR intervals (sample entropy [14]). They suggested that this HRC index could be used as an indicator of the risk level of developing adverse events including sepsis and death. Compared with infants with low-risk HRC index, infants with high-risk HRC indexes had 5- to 6-fold increased risk for an adverse event in the next day and 3-fold increased risk in the next week [15, 18]. A randomized trial enrolled 3,003 Very low birth weight (VLBW) infants across 9 NICUs conducted by Moorman et al. [22] showed that there was a trend toward increased days alive and ventilator-free for those infants whose HRC monitoring was displayed. A secondary analysis of clinical and HRC data from the same population further reported that the continuous HRC monitoring is associated with lower septicemia-associated mortality in VLBW infants, and this might be attributed to the earlier detection in the earlier course of sepsis [23]. Another retrospective case-control study [24] compared the highest HRC indexes in the 48 hours preceding blood culture sampling in LOS cases to the highest HRC indexes at the same postnatal days in controls, and found that an increase of HRC index >2 has a significant correlation with the diagnosis of LOS, supporting the utility of HRC monitoring to assist early detection of LOS.

Overall, in a series of studies on the usefulness of HRC analysis for early diagnosis of neonatal infections, most indicate that HRV analysis can diagnose sepsis 12 to 24 hours before traditional clinical methods, while one observed changes in HRC as early as 3 to 4 days before the onset of sepsis [14]. These studies reported satisfactory sensitivity for HRV analysis in the early diagnosis of sepsis, although specificity is somewhat compromised [25]. However, works from Griffin et al. have received marked criticism in the literature, particularly on the limited generalizability, over-fitting, lack of transparency in feature selection, limited consideration of clinical context (like gestational age, birth weight, and comorbidities), the lack of clear explanation for why certain features were important and, most importantly, the limited evaluation of model performance in real-world settings [26]. Furthermore, it remains unclear what exactly is being measured that far in advance (especially 3 to 4 days) of the clinical diagnosis of infection and whether altered HRV provides an early warning or an early detection of the presence of infection [25].

More recently, extended analyses of Heart rate variability, apart from the HRC, have been conducted in this field [27]. Classical HRV features from three domains are extracted and analyzed: time domain, frequency domain and the non-linear domain. Detailed calculations and indications of common HRV parameters are described in Section 2.2.5. Looking at specific Heart rate variability parameters, concerning time-domain indices, the septic neonates turn out to have lower mean

RR intervals [28], i.e., higher heart rate, and higher root-mean-square of successive RR interval differences (RMSSD) [29]. The percentage of decelerated RR intervals (pDec) decreases significantly in LOS infants in comparison with the controls prior to sepsis [30]. Frequency domain analysis is a powerful tool to examine the autonomic contribution to HRV. A decreased normalized HF (high-frequency power) was also found preceding the sepsis onsets [30]. In terms of the non-linear domain, it is reported that the lower approximate entropy, sample entropy and lower long-range fractal exponent ($\alpha 2$), i.e., uncorrelated randomness and complexity of the heart rate, are significantly associated with LOS sepsis [31, 32]. Greater SD2 than SD1 derived from Poincare analysis were also reported in extremely low birth weight septic neonates as compared with healthy neonates [28]. However, the acceleration capacity (AC) and sample asymmetry analysis (SAA) seem to have opposite indications in two different studies. In [33], the study setting was case-control states in septic infants, and both AC and SAA showed a decrease before the sepsis moments. While in [32], they used a case-control patient setting and differentiated the sepsis and non-sepsis at a patient level, and they observed that the early prediction of LOS was related to the increased SAA and AC. The visibility graph analysis of inter-beat time series has been also reported as a potential complementary tool for neonatal sepsis detection [34].

In conclusion, research during the last 20 years has shown that HRV analysis can be potentially useful for predicting LOS in preterm newborns. Although specific clinical features are still lacking, the underlying mechanisms associated with these findings are related to the lack of Autonomic nervous system development and cardiovascular instability in this population. However, serious limitations remain around this subject. Many studies on HRV-based prediction of LOS are based on small sample sizes, which can limit the generalizability of their findings [26]. Also, HRV signals are the results of a formally complex and multifactorial set of intertwined pathophysiological functions that are very difficult to separate. These aspects might explain the lack of repeatability and specificity in most of these studies and warrant further research in this field. We further analyze in this work the utility of HRV in this context, by taking into account the causal timeline of sepsis progression. The following section presents an original formalization of the causal events leading to the clinical detection of a sepsis event. This formalization has been performed in order to better apprehend the complexity of this problem and to propose novel analysis approaches.

5.1.3 Casual timeline of sepsis progression

The original diagram shown in Figure 5.1 presents a detailed timeline of the progression from the onset of infection to the clinical suspicion and confirmation of late-onset sepsis in neonates, built from an analysis of the literature and iterative discussion sessions with a NICU expert. It highlights the temporal relationship between physiological changes, such as early Autonomic nervous system (ANS) responses and Heart rate variability (HRV) alterations, alongside biochemical markers like C-reactive protein (CRP) elevation, all of which are critical in identifying the onset of infection and guiding timely clinical intervention.

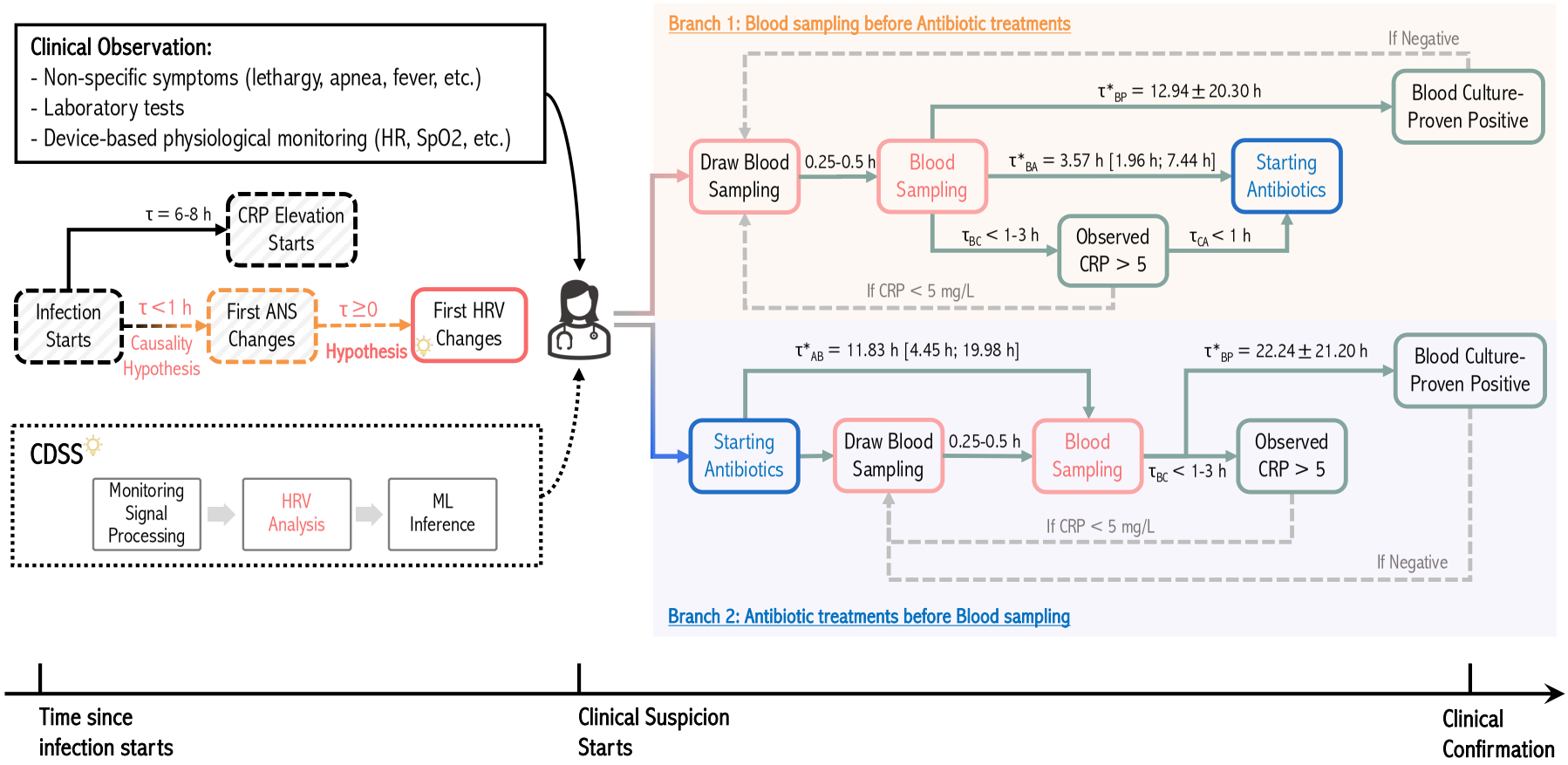

Figure 5.1: Original diagram formalizing the causal timeline of sepsis: From infection onset to clinical confirmation with positive blood culture. This diagram presents the timeline of key causal events in the progression of neonatal late-onset sepsis. It includes physiological events such as ANS and HRV changes, CRP elevation, and antibiotic administration. Transition durations were sourced from the literature, while the durations marked with "*" were calculated from the studied CARESS-Premi database. Dashed boxes are unobservable events.

Unfortunately, an early, direct observation of the onset of an infection event is not possible. In Figure 5.1, after the (unobservable) onset of infection, one of the first physiological effects is the response of the autonomic nervous system, typically occurring within the first hour, signaling the body's initial reaction to the infection. Approximately 6 to 8 hours later, CRP levels begin to rise, providing a widely recognized biochemical marker for inflammation and potential infection [35].

During NICU monitoring, clinical teams observe preterm infants for signs of sepsis using a combination of clinical assessments, laboratory tests, and device-based physiological monitoring. Based on these multiple and heterogeneous markers, a decision is taken between two scenarios, to either perform a blood sampling, in order to measure CRP and perform blood culture (branch 1 in Figure 5.1), or to directly apply antibiotics immediately without further evaluation (branch 2 in Figure 5.1). The time at which this decision is taken is typically used as the **timing for clinical suspicion for sepsis**. In the first scenario, blood sampling is conducted prior to the administration of antibiotics. CRP levels are then measured to confirm whether they exceed the critical threshold of 5 mg/L. If CRP $>$5 mg/L is observed, antibiotics are then administered, and this process usually takes 1 to 3 hours after blood sampling is documented. Summarized from the database used in this study, the overall time from clinical suspicion to confirmation of sepsis through CRP is approximately 3.57 hours, with a final confirmation based on blood culture results occurring within 12.94 hours. In the second scenario, antibiotics are administered before blood sampling. Blood is drawn afterward to measure CRP levels and confirm the presence of infection. The time from the start of antibiotics to the final confirmation of sepsis is approximately 22.24 hours, based on the available dataset. In some cases, even if a blood sample is taken, no significant CRP elevation is observed. The clinical surveillance continues and further blood sampling might be performed afterwards.

As already discussed, there is thus currently a major clinical need to provide an indirect, early and robust marker of infection, and use it in a clinical decision support system (CDSS) to assist clinicians in the monitoring and therapeutic strategies. One of the major hypotheses in this field is that the early autonomic alternations described above might be manifested by changes in HRV and that the analysis of HRV might be thus used as an observation window to the onset of sepsis. The attempts of building CDSS based on HRV analysis are built on this hypothesis. However, many factors explain the current lack of performance and generalization of current solutions:

- Limited understanding of early signs of infection:
  Although the early autonomic response to infection is a well-accepted hypothesis, many other physiological factors are involved. Research is currently very active on this subject.
- Variability in clinical practice:
  Clinicians may have different practices and protocols for monitoring preterm babies, making it difficult to homogenize multi-centric data collections and annotations.
- Lack of standardized definitions:
  There is no universally accepted definition of neonatal infection.
- Complexity of LOS diagnosis:

As described in Figure 5.1, infection diagnosis requires considering multiple factors including clinical signs, laboratory test results, and physiological monitoring data. These factors are related to different underlying physiological processes and occur at different timings or even different time scales.

- High false positive rate:
  The risk of false positives (i.e., detecting infection when it is not present) is particularly high due to the complexity of the diagnosis and the variability in clinical practice.

As a consequence of the factors listed above, a precise definition of infection and, therefore, a precise definition of the **beginning of an infection event** is thus a mathematically formal complex task, without any concrete solution today. Since the definition of these annotations (infection state or timing of beginning of infection) will be used as the main target to create ML approaches, supervised model training in this field suffers from significant sources of bias. Indeed, many sepsis prediction models perform well when trained with treatment-related information, but they often fail to provide clinicians with new insights, as much of their predictive power comes from the clinical suspicion itself, which already plays a major role in distinguishing between patients' health states [36].

In this study, we defined the clinical suspicion time by whichever occurs first, either antibiotic administration or CRP>5 mg/L. This approach helps minimize the risk of "label leakage" when annotating the data. Furthermore, a significant effort was directed to perform a multi-expert clinical event classification and annotation of each event, as well as the translation of each of the annotated events into a detailed labeling process, used for ML training. These crucial points are described in the following sections.

5.1.4 Proposed approach

In this study, we aim to develop a non-invasive, computer-assisted diagnostic tool for early detecting LOS in preterm infants, leveraging continuous ECG data routinely collected in NICU. By applying advanced signal processing techniques and machine learning algorithms to real-life monitoring data, we seek to identify specific HRV patterns that precede the clinical diagnosis of sepsis. The study also aims to assess the effectiveness of HRV analysis derived from real-time monitoring in enhancing diagnostic accuracy and timeliness, ultimately contributing to more informed and timely clinical interventions.

In the following sections, we first describe the clinical trial used in the study, including the study population and the clinical classification of sepsis-related events. Then detailed strategies for constructing the database are presented. This includes the processing chain from the acquisition of ECG signals to the computation of HRV features and the implementation of a novel labeling strategy. We then outline the methodology for developing and evaluating machine learning models for sepsis detection, which involves generating multiple variants of the feature sets and ML

models. Then, we present the performance of all proposed ML models in detecting sepsis, as well as specific HRV patterns observed around the onset of sepsis in our dataset. Finally, we provide a comprehensive discussion of our results, situating them within the broader context of existing literature and exploring their wider implications.

5.2 Study Population and Clinical Events Classification

5.2.1 Study population (Patient-level)

The data used in this study is derived from a prospective multi-center clinical study CARESS-Premi, as introduced in Section 2.1.1. The CARESS-Premi database collected continuous cardio-respiratory monitoring signals, general clinical records and available biological tests during enrolled infants' hospitalization. This study considered the entire database and included all eligible patients at the three clinical centers (Rennes, Lille and Angers).

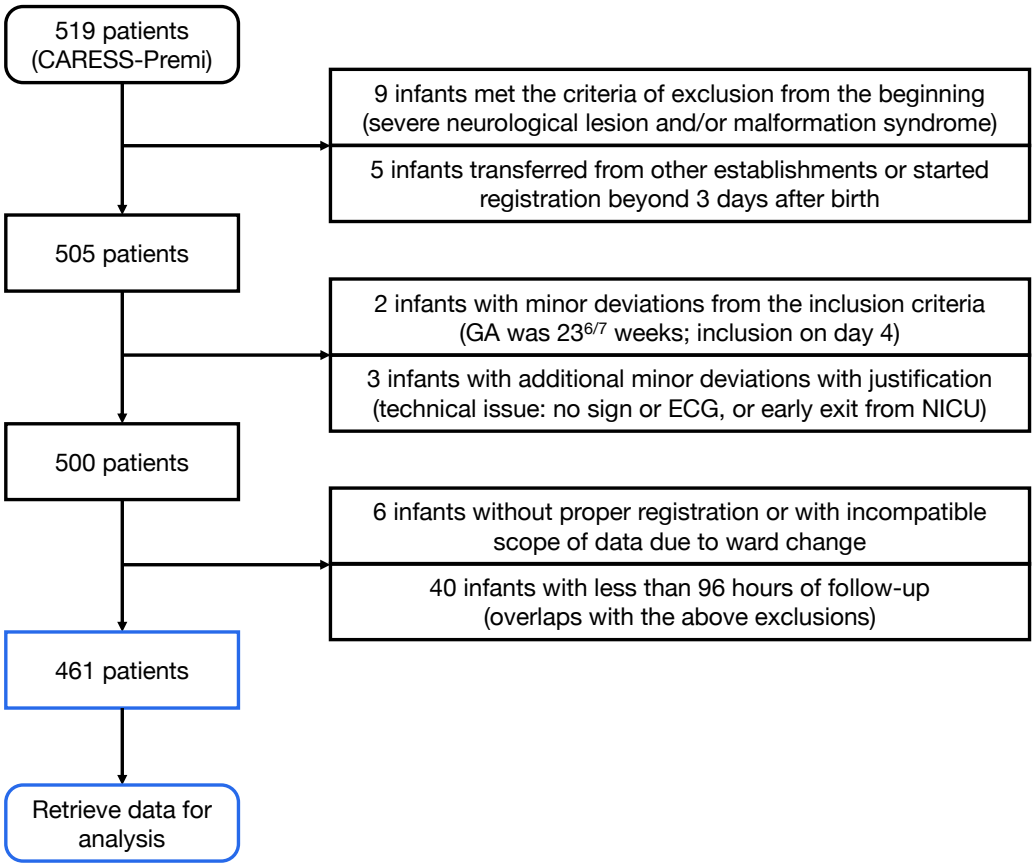

Figure 5.2: Study population inclusion and exclusion for late-onset neonatal sepsis.

Within the CARESS-Premi cohort, some patients who met the non-inclusion criteria and/or encountered deviations were excluded from the study, including infants with neurological impair-

ments, congenital malformations, administrative issues, technical issues with data collection after day 3, or who were lost to follow-up, etc. Besides, infants with a confirmed and only episode of sepsis within 72 hours after birth were considered to have early-onset sepsis and were therefore excluded, as were infants with less than 96 hours of follow-up. Others were excluded from the study due to short follow-ups. As a result, 461 infants remained for analysis. Figure 5.2 shows the flow chart of study population inclusion and exclusion.

5.2.2 Multi-expert clinical event classification (Event-level)

Clinical classification of Late-onset sepsis (LOS) events was in accordance with the NEO-KISS (Neonatal Nosocomial Infection Surveillance System) protocol [37] for nosocomial infection surveillance for preterm infants was employed. To confirm or reject suspicious events as sepsis episodes, a multi-expert blind analysis and consensus classification approach was conducted. It involved an initial blind analysis by three neonatologists, followed by a collaborative consensus-seeking procedure and, if necessary, further review of patient health records to resolve disagreements. Finally, all suspected sepsis events were classified into 8 sub-categories in total.

In this study, we combined "clinical sepsis", "pneumonia", "urinary tract infection", "enterocolitis", "laboratory-confirmed bloodstream infection (LCBI)" and "LCBI with Coagulase Negative Staphylococci (CNS) as the sole pathogen", and qualified them as positive events (EVT_pos) of late-onset neonatal sepsis.

In contrast, suspected events classified by the multi-expert blind analysis as "isolated inflammation" and "no systemic infection, no inflammation" were considered negative events (EVT_neg) that should be excluded from the study since these episodes may have been contaminated by infection but could not be identified as sepsis.

In the CARESS-Premi cohort, 454 suspicious events from 287 patients were identified and documented. After consensus reached by multiple experts, the categories and number of the sepsis-relevant events are as follows:

- EVT_pos (confirmed sepsis events):
  + Clinical Sepsis (172 events)
  + Pneumonia (24 events)
  + Urinary Tract Infection (16 events)
  + Enterocolitis (33 events)
  + Laboratory-Confirmed Bloodstream Infection (LCBI) (14 events)
  + LCBI with Coagulase Negative Staphylococci (CNS) as the sole pathogen (40 events)
- EVT_neg (denied/rejected sepsis events):
  - Isolated Inflammation (85 events)
  - No Systemic Infection, No Inflammation (70 events)

To simplify terminology in this study, we refer to septic patients as those diagnosed with at least one confirmed sepsis event during the follow-up period; whereas non-septic patients were those who never had any confirmed sepsis events.

Note that this more formal, standardized and thorough definition of sepsis events is not the same definition used in previous works by our team [34].

5.3 Dataset Construction

Figure 5.3 illustrates the workflow of constructing the dataset for the study. It includes three branches: signal processing, clinical data matching (metadata), and label generation.

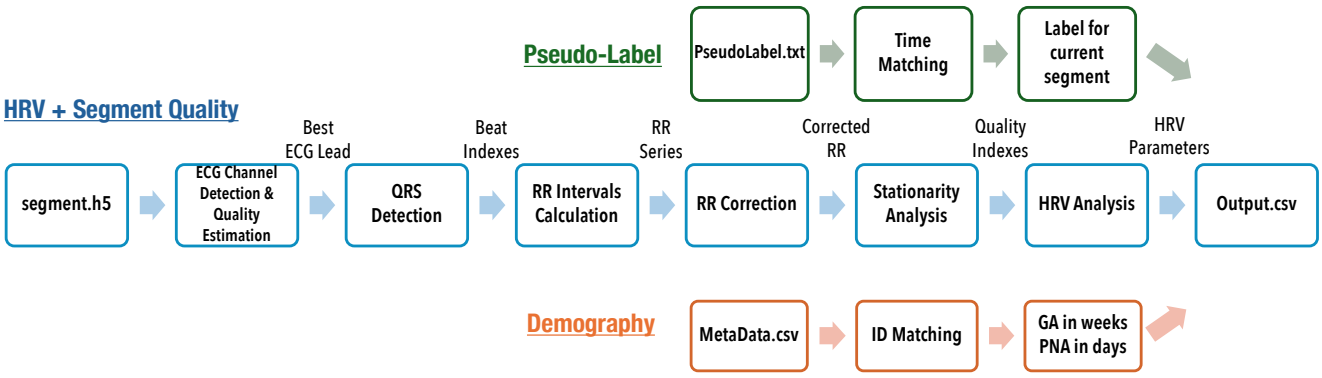

Figure 5.3: Illustration of data preparation for early sepsis detection: signal, metadata, pseudo-label.

5.3.1 Signal preparation and data processing (Sample-level)

With the benefits of the ASCENT system (details refer to Section 2.1.2) proposed by our team, we retrieved all available monitoring signals and clinical records during the follow-up of each eligible infant for this study.

For all concerned infants, signals were queried and compiled in consecutive, non-overlapping 6-hour segments from postnatal day 3 to the end of follow-up. This was done by specifying a metadata file containing timestamps for each patient at 6-hour intervals from day 3 of birth to the end of follow-up as anchor points. A detailed description of data extraction and compilation refers to Section 2.1.3. The compiled data in HDF5 format has three types of electrophysiological, while in this study, we focus on only ECG signals for extracting interested HRV parameters.

Regarding ECG signal processing and feature extraction, we employed the proposed pipeline

for real-life monitoring signals, as thoroughly presented in Section 2.2. In brief, it can be divided into three steps: *1)* signal pre-processing for raw ECG signals, *2)* RR series calculation and correction, and *3)* HRV analysis.

The raw 3-lead ECG signals were obtained with a sampling frequency of 500 Hz. We first implemented algorithms that detect all valid channels for each segment and remove the noisy parts. The rolling standard deviations of the normalized signals were then calculated to assess the stability of the time series, which was used as a measure of signal quality. The best denoised ECG channel was then chosen for further analysis. See detailed approaches in Section 2.2.1.

An improved QRS detector with filter coefficients specifically adapted for newborns, as proposed in [38], was utilized to identify the R-peaks from ECG waveforms, followed by simple calculation of difference to obtain RR intervals (refer to Section 2.2.2). Next, we developed a rule-based RR correction algorithm to match and correct the R-peaks that were wrongly detected by the QRS detector, especially in cases where the waveforms were noisy and complicated (refer to Section 2.2.3).

From the corrected RR series, we calculated all the HRV parameters described in Section 2.2.5. The time domain analysis consists of the extraction of the mean (Mean), the standard deviation (Std), the median (Median), the skewness (Skewness) and the kurtosis (Kurtosis), and the inter-decile range (IDR) of the RR series, which provide an initial indication of the global variability. Another common feature, the square root of the mean squared differences of the successive RR intervals (Rmssd), was also computed, and it reflects short-term beat-to-beat variations and quantifies more specifically the parasympathetic modulation of the autonomic nervous system [39]. Acceleration capacity (AC) and deceleration capacity (DC) were included, and they are based on the phase-rectified signal averaging method which is much more robust to non-stationarity [40]. Furthermore, we extracted some features specifically designed to capture neonatal HRV, such as sample asymmetry of the RR intervals histogram (SampAsy) [14], percentage of the RR intervals longer than the mean RR of the previous certain intervals (pDec) and its standard deviation (stdDec) [41].

The frequency domain analysis estimated the power spectrum integrating the low-frequency (LF: 0.02-0.2 Hz) and high-frequency band (HF: 0.2-2 Hz) obtained by autoregressive modeling of the 4 Hz resampled RR series. The normalized units of the features (LFnu and HFnu) as well as the ratio of LF and HF power (LFHF) that reflects sympatho-vagal balance were also calculated.

Regarding non-linear measurements of HRV, sample entropy (SampEn), configured with a window length of 3 intervals and a tolerance of 0.25, was calculated to quantify the regularity and predictability of the given RR series [12]. Poincaré plot analysis [42] was performed to capture the short- and long-term variability represented by the width (SD1) and length (SD2) of the ellipse in the plot. Another two coefficients derived from a self-similarity parameter that characterizes the

long-range fractal correlation properties of the signal were obtained through detrended fluctuation analysis [43]. We evaluated the fractal scaling exponent from 4 to 40 beats ($\alpha 1$) and from 40 to 1000 beats ($\alpha 2$) [44].

Accordingly, we retrieved all the raw data of the 461 eligible patients available in the database and compiled them in 6-hour segments, the data points at time t represent the 6-hour epoch after, i.e., [t, t+6] hours. Of these, monitoring signals of some patients were not properly acquired and/or uploaded before the stabilization of the ASCENT system (refer to Section 2.1.2), resulting in a certain loss on the valid population for the study. In addition, during the employment of the proposed signal processing pipeline, some segments or patients were rejected by our data processing chain due to poor data quality, in which a set of strategies were integrated to guarantee the Signal-to-noise ratio (SNR). These screenings left 400 infants remaining. On the other hand, there were also some data segments that were not able to be labeled due to missing clinical records. It should be underlined that, although at the patient level, some patients were eligible for further analysis, from the perspective of each patient's follow-up timeline, not all data segments from day 3 to the end of follow-up were successfully processed and thus used to constitute the final data set, since some data segments may be removed due to unavailability of the original signals or due to low SNR during processing, resulting in data discontinuity in the timeline.

Meanwhile, in addition to HRV features, demographic information such as Gestational age (GA) and Postnatal age (PNA) from the Electronic health records (EHR) (MetaVision, iMDsoft, Tel Aviv, Israel) were collected as part of the feature set as they proved critical determinants of the neonatal host response to sepsis [45].

The branch of creating pseudo-labels is detailed in the following section.

5.3.2 Pseudo-labeling strategy

Considering the six-hour granularity of the retrieved raw ECG and the derived HRV characteristics, we proposed a logical rule-based labeling strategy to generate corresponding pseudo labels for data samples on a six-hour basis. The labels are divided into three categories:

- "**1**" represents positive samples, used for samples in septic condition;
- "**0**" represents negative samples, used for non-septic conditions;
- "**-1**" corresponds to samples in which the infant's condition is **uncertain**, and thus these samples will be discarded and used **neither as cases nor as controls**.

As the clinical classifications (Section 5.2.2) only specified the dates of suspected events (EVT_pos and EVT_neg), we should further identify detailed time for the following labeling process. So, the events were first identified by the date of their occurrence, as only one event could occur on any given day. Then, the specific time of the events could be precisely defined by the date and time of antibiotic administration or blood culture and C-reactive protein (CRP) collection, depending on

the purpose. Based on clinical perspectives, we defined the specific **onset time of sepsis events** as the earliest among antibiotic administration, CRP results higher than 5 mg/L and positive blood cultures (often recorded as the time performing blood sampling).

To better generate the labels, the clinical records in the CARESS-Premi database were queried, cleaned, merged, and screened at the patient level to identify the starting and ending dates and times (if present) of relevant events. These events include follow-up, antibiotic treatment (ATB), CRP higher than 5 mg/L (CRP5), positive blood cultures (posBC), and the classified infection events with clinical annotations (EVT_pos and EVT_neg, as described in Section 5.2.2). Missing dates or times were imputed by reasonable inference based on the clinical records, for instance, the missing end dates of CRP5 were imputed by the dates when the maximum CRP results were measured. Personalized event timelines including interventions and clinical findings mentioned above were generated.

The labeling strategy was formalized as follows, where "|", "!" and "&" denote *or*, *not* and *and*, respectively:

- Label $= 1$: EVT_pos | (! EVT_neg & ATB & CRP5)
- Label $= -1$: (! EVT_pos & EVT_neg) | (! EVT_pos & ! EVT_neg & ((ATB & ! CRP5) | (! ATB & CRP5)))
- Label $= 0$: Others.

We considered a 6-hour data segment to be positive ("1") when it was on the day been annotated as a confirmed sepsis event (EVT_pos), or when both the ATB and CRP5 were registered and at the same time it was not on the same day of an annotated denied sepsis. Uncertain samples ("-1") will be excluded from the analysis because they have neither been diagnosed/classified with sepsis nor could be used as control samples due to the presence of certain suspected symptoms. The data segments in two conditions were labeled as "-1": when the segments were on the day when it was classified as denied sepsis events (EVT_neg); and when antibiotic treatment (ATB) and the CRP5 were not administered simultaneously. Samples in the remaining conditions were labeled as "0" and used as control samples.

Shown in Figure 5.4 and Figure 5.5 are some examples of the labeling results. This strategy integrates the neonatal caregivers' clinical expertise into the labeling procedure, i.e., the documented moments of their initial clinical suspicions and immediate interventions. When they have an early and accurate "gut feeling" of the babies developing sepsis, our corresponding labels will be equally timely and potentially predictive.

5.3.3 Post-processing on the labels

The proposed labeling strategy was designed to take full account of the onsets and ends of suspicious septic events. The logical rules work in a *in-place* manner, meaning that the label of each 6-hour segment is determined solely by the values of the corresponding elements at the same

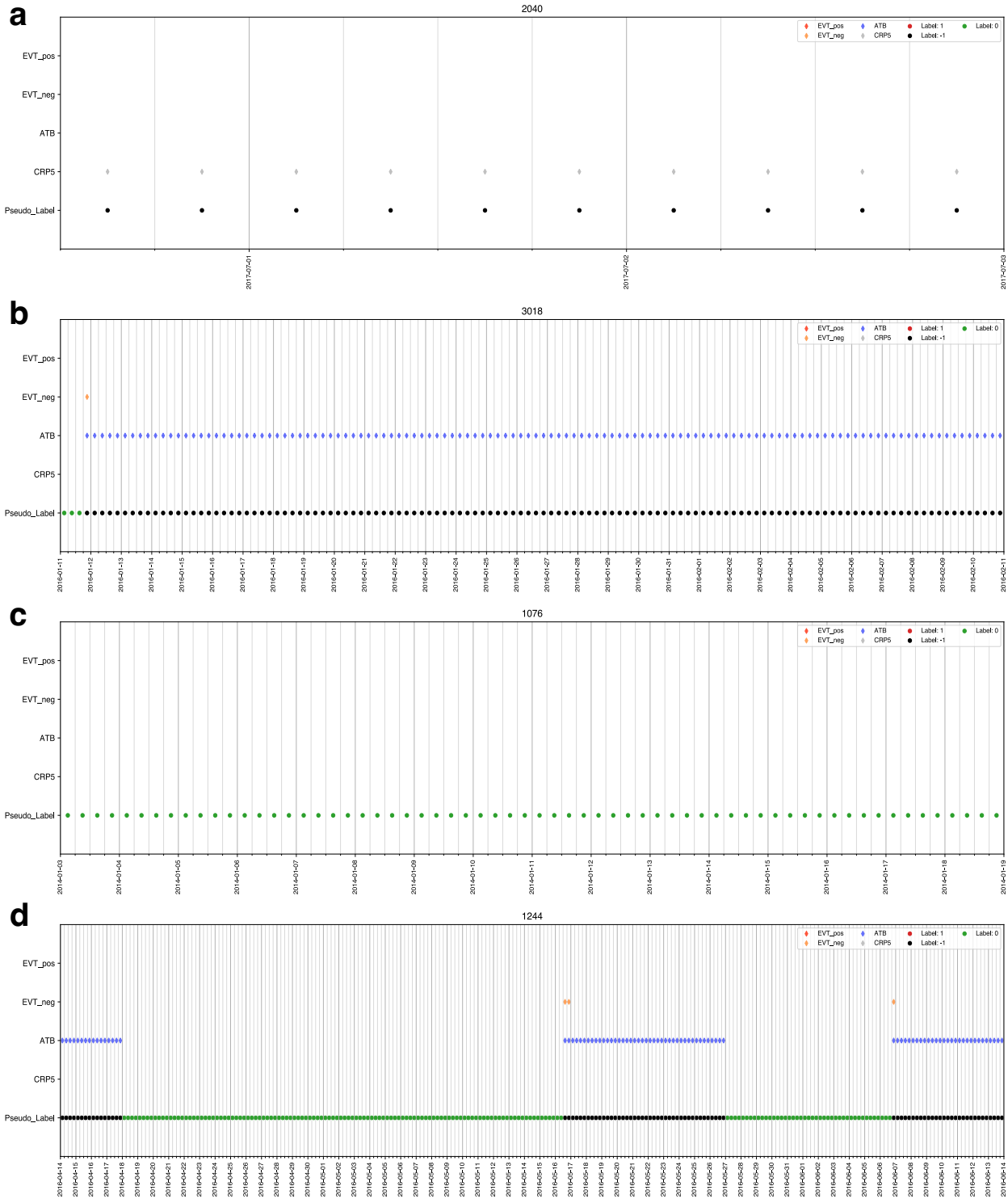

Figure 5.4: Visualization of the proposed pseudo-labeling strategy (1/2). (**a**) Patient 2040 was labeled as all "-1". (**b**) Patient 3018 had one suspicious event but was denied having sepsis (EVT_neg) and the associated segments were labeled as "-1". (**c**) Patient 1076 had no suspicious events during the follow-up and all data segments were labeled as "0" (controls). (**d**) Patient 1244 had two suspicious events but was denied having sepsis because of the absent CRP5 during the antibiotic treatments. Each major vertical grid denotes 24 hours, in which each minor grid denotes a 6-hour epoch composited by [t, t+6] hours.

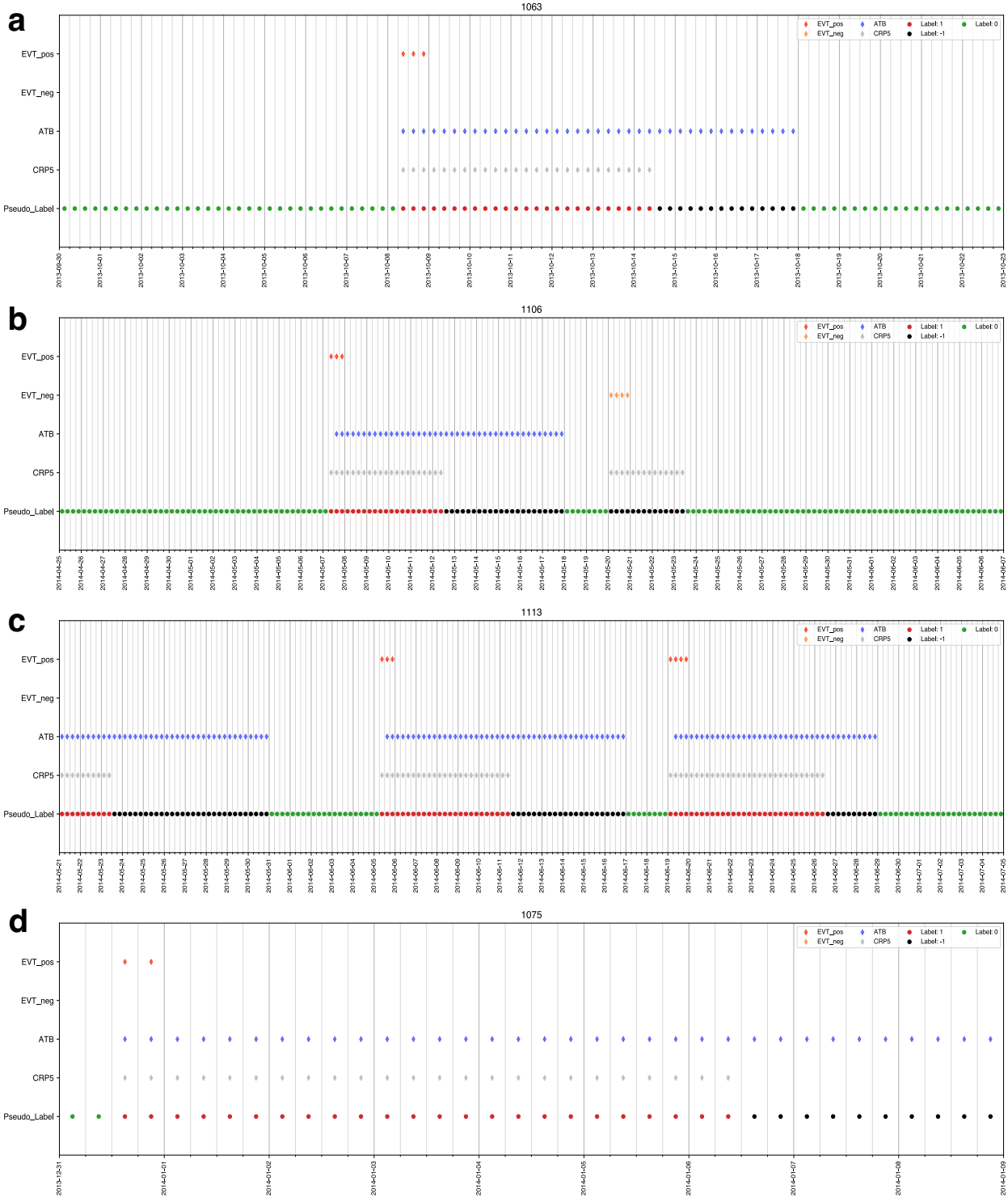

Figure 5.5: Visualization of the proposed pseudo-labeling strategy (2/2). (**a**) Patient 1063 had one confirmed septic event (EVT_pos) and had three types of labels during follow-up. (**b**) Patient 1106 had one confirmed sepsis and one rejected sepsis. (**c**) Patient 1113 had two confirmed septic events. (**d**) Patient 1075 had a relatively shorter monitoring duration with one confirmed sepsis event at the beginning of follow-up. Each major vertical grid denotes 24 hours, in which each minor grid denotes a 6-hour epoch composited by [t, t+6] hours.

index (time segment) in EVT_pos, EVT_neg, ATB and CRP5 without knowing any time-relevant information, such as the labels of previous or subsequent data segments.

However, in practice, the pre-defined granularity of signal segments was not that fine (6 hours) and the inconsistently recorded resolution of different clinical events, for instance, the suspicious septic events were annotated with dates while the ATB and CPR5 were recorded with specific dates and times. These factors may lead to wrong labeling as the strategy focuses on *local* information without a *global* perspective.

By observation, the post-processing considered two patterns of mislabeling:

- Mislabeling "-1" as "1" with a pattern as: -1 -1 -1 -1 -1 -1 -1 1 1 -1 -1 -1 -1 -1 -1 -1
  This usually happens when both ATB and CRP5 are documented, but the associated event is retrospectively classified as non-sepsis by experts.
- Mislabeling "1" as "-1" with a pattern as: 1 1 1 1 1 1 1 -1 -1 1 1 1 1 1 1 1
  This often occurs when a confirmed sepsis event is accompanied by asynchronous antibiotic treatment (ATB) and high CRP results (CRP5), or when there is a short absence of ATB or CRP5 between consecutive confirmed sepsis events.

Figure 5.6 shows examples of the mislabeling of the pseudo-labeling strategy and signifies the necessity to perform post-processing to correct the labels to the greatest extent. According to the logical rules of the labeling strategy, the first pointed red sequence of the 2 data segments was labeled as "1" (in red) because of the overlap between the clinical events of ATB and CRP5. Yet, from a clinical perspective, these two segments should be labeled as "-1" (in black) since it was part of a negative sepsis event that was suspected on 2013-09-18. On the contrary, the second pointed sequence of 6 data segments was labeled as "-1" (in black) due to the delay in ATB administration relative to the positive results of CRP5, and it should be labeled as "1" (in red) as they were both associated with the sepsis event on 2013-09-19. Besides, the third segment of 3 black dots is a special case that can be assigned as both black ("-1") or red ("1").

Therefore, we implemented a simple set of rules that match the two mislabeling patterns mentioned above to locate the wrong labels and then correct them. This was conducted in a patient-by-patient and event-by-event manner to post-process pseudo-labels from a more comprehensive time-series perspective.

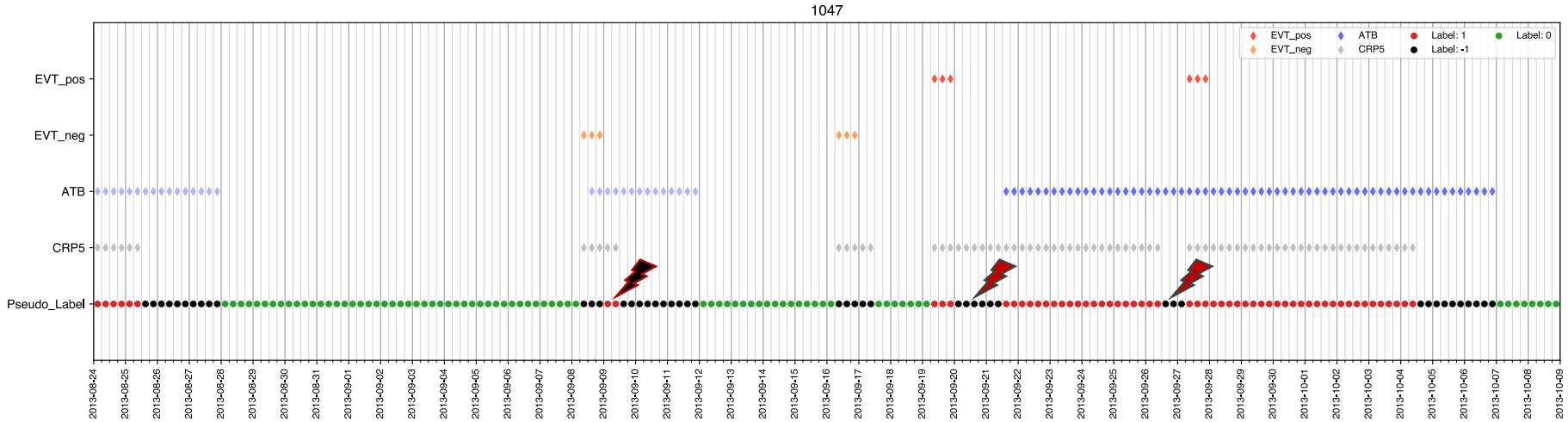

Figure 5.6: Examples of the limitations in the pseudo-labeling strategy.
The first pointed sequence should be labeled as "black"; the second pointed sequence should be labeled as "red"; and the third pointed sequence can be both "black" or "red". Each major vertical grid denotes 24 hours, in which each minor grid denotes a 6-hour epoch composited by [t, t+6] hours.

5.3.4 Dataset overview

In brief, the final composition of the constructed dataset depends on the intersection of data availability at three levels: patient-level (Section 5.2.1), event-level (Section 5.2.2) and sample-level (Section 5.3.1), as revealed in Figure 5.7. Through all steps in the dataset construction process, we obtained the processed features with pseudo labels for 399 infants, accounting for 189 confirmed sepsis episodes and 33,355 samples (each represents one 6-hour segment).

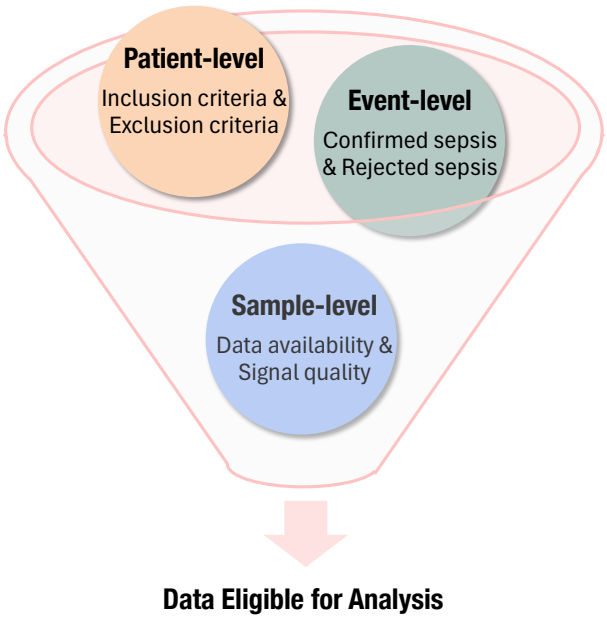

Figure 5.7: Dataset construction based on data availability at the patient, event and sample levels.

At the patient level, as shown in Figure 5.8, 5 patients (in gray) had all valid data segments labeled as "-1" that should be removed from the analysis because the samples cannot be considered as positive (sepsis) nor negative (non-sepsis); 243 patients (in green) can be regarded as non-septic patients (controls) because all their data segments were labeled as "0". A total of 151 patients were identified as septic patients as they had data segments labeled as "1", and it consisted of 143 patients with both mixed labels, 3 patients with only positive labels and 5 without uncertain labels. Altogether, 394 patients were eventually included for analysis.

At the event level, different from the common way performed in the previous studies that only included the first proven Late-onset sepsis episodes [29, 32, 34], all confirmed LOS events were considered in this study. Within 189 included septic patients, the mean number of confirmed sepsis events was 1.25, ranging from 1 to 3.

From the sample level, where one sample refers to the extracted features with a pseudo label derived from one 6-hour segment of raw ECG, a total of 33,355 samples were obtained from the included 394 patients. After removing the unconfirmed samples labeled as "-1" (6,562 samples),

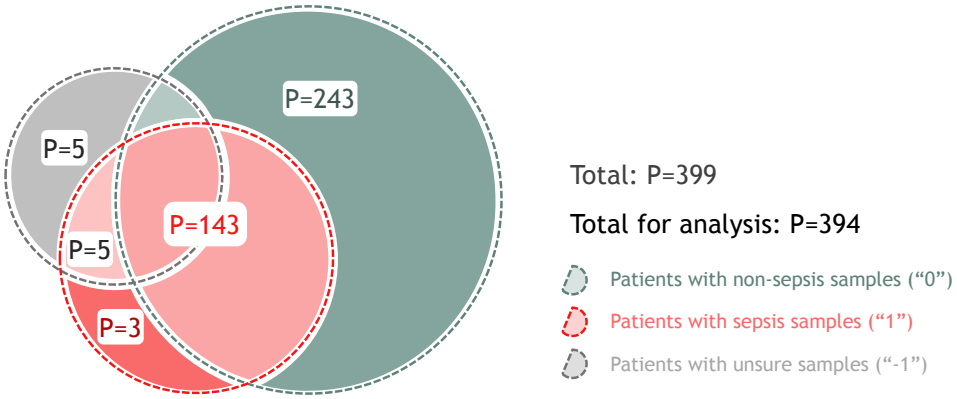

Figure 5.8: Composition of the included population: at the patient level.

there were 3,169 positive samples ("1") and 23,624 control samples ("0"), with a class ratio of 1:7.5 in general. The median number of available samples per patient is 59 (ranging from 3 to 215 samples) equivalent to 354 hours (15 days) of monitoring segments.

5.4 Machine Learning Models Development and Evaluation

5.4.1 Overall strategy of ML models training and evaluation

Aligning the extracted features (Section 5.3.1) and the corresponding labels (Section 5.3.2) in time constituted the dataset for supervised learning of classification. We adopted the Monte Carlo cross-validation (refer to Section 2.4.3 for more information) to develop and evaluate the ML models. We designed a two-stage Monte Carlo experimental procedure containing the "preliminary" Monte Carlo first to find the optimal set of hyper-parameters for machine learning algorithms and the "ultimate" one to train a set of models with the optimized hyper-parameters and then estimate the performance and generalizability of the proposed models.

As shown in Figure 5.9, the dataset was first split into a train set (80%) and test set (20%) on the patient level, and a random down-sampling was conducted to the majority class (class labeled "**0**") for balancing the class ratio to 1:1. Next, two branches were depending on the stages, for preliminary stage a grid search within a 5-fold cross-validation framework was performed on the balanced train data for optimizing the hyper-parameters of models; while for the ultimate stage the hyper-parameters for building models were fixed to the optimized ones. In both cases, models were then fitted on the whole balanced train set and validated on the left-out test set.

5.4.2 Machine learning algorithms

The detection of neonatal late-onset sepsis in this study was implemented as a binary classification task, distinguishing each data segment as either septic ("1") or non-septic ("0") using ex-

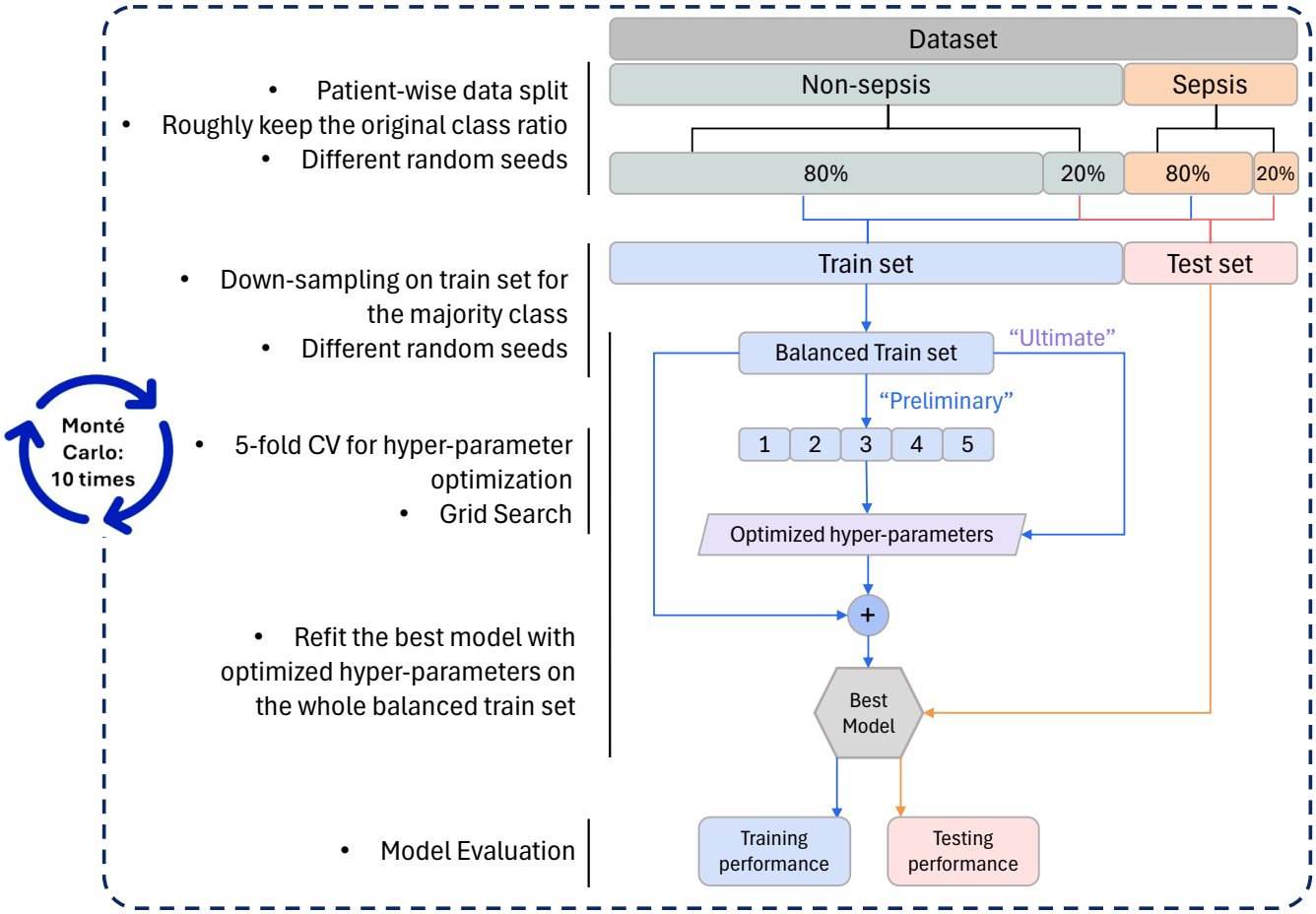

Figure 5.9: Two-stage Monte Carlo procedure for hyper-parameters optimization, ML model training and evaluation.

tracted HRV features and basic clinical information. A series of classic supervised machine learning algorithms for classification were employed, including Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Multilayer perceptron (MLP). Refer to Section 2.4 for more information. Additionally, our team has long-term (more than 20 years) experience in the design, implementation and application of recurrent or convolutional neural networks for the analysis of time series [46, 47]. Thus, we designed a shallow Convolutional neural networks (CNN) with special structures to achieve sepsis detection in our case.

Based on the length and the eventual phase distortion of the original data used for model training and evaluation, these algorithms can be divided into instantaneous (time-independent) detectors and time-dependent detectors.

Time-independent (Instantaneous) detectors

The instantaneous detectors indicate that when deciding whether the current segment of data

(s) is labeled as "1" or "0", only the current information (HRV parameters, GA and PNA derived from the segment in question) is used for analysis, formulated as Equation 5.1, where s denotes the index of data segment, X denotes the HRV features, $\mathcal{M}(\cdot)$ stands for models and $\widehat{L}$ are predicted labels. In this case, the samples are independent of time, and randomly disrupting the order of the samples will not affect the training of the classifier model. So there is no past or future information involved when making decisions, which enables the on-the-edge deployment and inference in a pseudo-real-time manner. Most detectors in the literature in this field are defined in this manner.

$$\begin{aligned} X(s) &= HRV(s) \\ \widehat{L}(s) &= \mathcal{M}(X(s) + GA(s) + PNA(s)),\ s > 0 \end{aligned} \tag{5.1}$$

Time-dependent detectors

We use time-dependent detectors to refer to models that require historical data to determine the class of the current segment. Specifically, there are several variants according to how the features are transformed and combined: $\Delta X(n)$ and $ConcatX(n)$.

The $\Delta X(n)$ represents calculating the differences between the HRV features of the current segment and of the previous segment with a step of n, as formulated in Equation 5.2a. We computed $\Delta X(1)$ and $\Delta X(2)$ in this study. Besides, another variant $\Delta refX(r)$ is to use the first (earliest) sample labeled "0" for each patient as a reference to calculate the difference, so the difference reflects the relative HRV characteristics of the current state compared to the control state. As we are using longitudinal and continuously monitored data, the time span of the reference sample r could be very different from segment to segment as well as from patient to patient. Equation 5.2b presents the formulation. It should be noted that the differencing operation Δ does not change the feature dimension. Note that this type of approach was used in one of the previous works from our team [34].

Another type of variant is $ConcatX(n)$, which concatenates the HRV features of preceding successive segments with a length of n, leading to an increase in feature dimension by n times (Equation 5.2c). We limited the length n to less than 6, which is equivalent to using at most 36 hours in the past to make classification.

$$\begin{aligned} \Delta X(n) &= HRV(s) - HRV(s-n),\ n \in \{1,2\},\ s > n \\ \widehat{L}(s) &= \mathcal{M}(\Delta X(n) + GA(s) + PNA(s)),\ s > n \end{aligned} \tag{5.2a}$$

$$\begin{aligned} \Delta refX(r) &= HRV(s) - HRV(r),\ r = min\{t \in [0,s) \mid L(t) \neq 1\} \\ \widehat{L}(s) &= \mathcal{M}(\Delta refX(r) + GA(s) + PNA(s)),\ s > r \end{aligned} \tag{5.2b}$$

$$
\begin{aligned}
ConcatX(n) &= [HRV(s), ..., HRV(s-n)],\ 0 < n < 6 \\
GA(n) &= [GA(s), ..., GA(s-n)],\ 0 < n < 6 \\
PNA(n) &= [PNA(s), ..., PNA(s-n)],\ 0 < n < 6 \\
\widehat{L}(s) &= \mathcal{M}(ConcatX(n) + GA(n) + PNA(n)),\ 0 < n < 6
\end{aligned}
\tag{5.2c}
$$

In this study, we target detecting the onsets and the presence of sepsis events by developing models that can discriminate the septic status from non-septic status. Thus, for the labels, we continue to use the original label of the current segment as the label for the new HRV variants. But, some constraints are proposed. Each sample (6-hour data segment) has its original label ("1" for sepsis and "0" for non-sepsis), and different state transition patterns will inevitably occur when multiple data segments are combined. There are four transition patterns from the first (the earliest) to the last (the current) segment in the context of a variant: "0" to "0", "0" to "1", "1" to "0" and "1" to "1". Specifically, the "0"–"0" pattern represents a stable non-septic state and the associated samples can be used as the negative class for training a classifier, and patterns "0"–"1" and "1"–"1" show a transition from non-sepsis to sepsis and a continuation of sepsis condition, suggesting the onset and presence of sepsis events, respectively, these samples are thus regarded as the positive class. However, the state transition from "1" to "0" indicates an end of a sepsis episode, However, a state transition from "1" to "0" indicates the end of the sepsis episode, it consists of implicit information about septic states while it is labeled as "0" (using the original label of the current sample). Using this sample to train a classifier will completely confuse the model's distinction between sepsis and non-sepsis. Therefore, data with this transition pattern were excluded.

	HRV Variant, Label	**Time-dependent Transition Patterns: (0 to 0) & (0 to 1) & (1 to 1)**																			
Eq.5.1	X(s)=HRV(s), L(s)	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	…
Eq. 5.2a	ΔX(1)=HRV(s)-HRV(s-1), L(s)	/	~~1~~	~~1~~	~~0~~	00	00	00	00	00	00	01	11	11	11	11	11	~~0~~	00	00	…
Eq. 5.2a	ΔX(2)=HRV(s)-HRV(s-2), L(s)	/	/	~~1~~	~~0~~	~~0~~	00	00	00	00	00	01	01	11	11	11	11	~~0~~	~~0~~	00	…
Eq. 5.2b	ΔrefX=HRV(s)-HRV(r), L(s)	/	/	/	/	00	00	00	00	00	00	01	01	01	01	01	01	00	00	00	…
Eq. 5.2c	ConcatX(3)=[HRV(s), HRV(s-1), HRV(s-2)], L(s)	/	/	111	~~110~~	100	000	000	000	000	000	001	011	111	111	111	111	~~110~~	~~110~~	000	…

Figure 5.10: Illustration of the transition patterns and labels considered in the time-dependent models trained using historical data. The first two columns represent the corresponding feature manipulation equations; the next colored grids (red: labeled as "1", green: labeled as "0", gray: excluded patterns) represent the transition patterns when performing different feature transformations.

Figure 5.10 illustrates examples of the generation of HRV variants. By manipulating the extracted HRV features, including performing differences between the current segments and the previous segments ($\Delta X(1)$, $\Delta X(2)$ and $\Delta refX$) and concatenating with the successive segments ($ConcatX(3)$, taking $n = 3$ as an instance), we obtained various values and dimensions of the HRV

features. In all cases, the labels of the newly generated samples are the original labels of the current segment (L(s)).

In addition to classical machine learning approaches, we proposed shallow CNN for LOS detection in order to take advantage of the convolution operator along our time series-derived structured data.

In our shallow CNN, unlike usually applying 1-dimensional convolution operations for time-series data, we instead regarded our data as "images" and employed 2-dimensional convolution operators in order to better explore the implicit information between successive data segments rather than in feature spaces. To achieve this, imitating a common CNN that takes 3-dimensional tensors as inputs (the height, width and the number of channels of an image), we constructed data matrices by considering the time axis as an extra dimension and shifting over the original feature vectors through a sliding window, equivalent to the $ConcatX(n)$ operation as described in Equation 5.2c. Then we composed the required three dimensions of input tensors by the number of features, the length of the concatenating window (3 or 6) and the channel set to 1.

As for the kernels in both convolutional and pooling layers, they were adapted as 1-dimensional kernels to fully exploit the information in the time axis without mixing up between features. Figure 5.11 and Figure 5.12 demonstrate the architectures of the proposed shallow CNN with 1-layer convolution and 2-layer convolution, respectively.

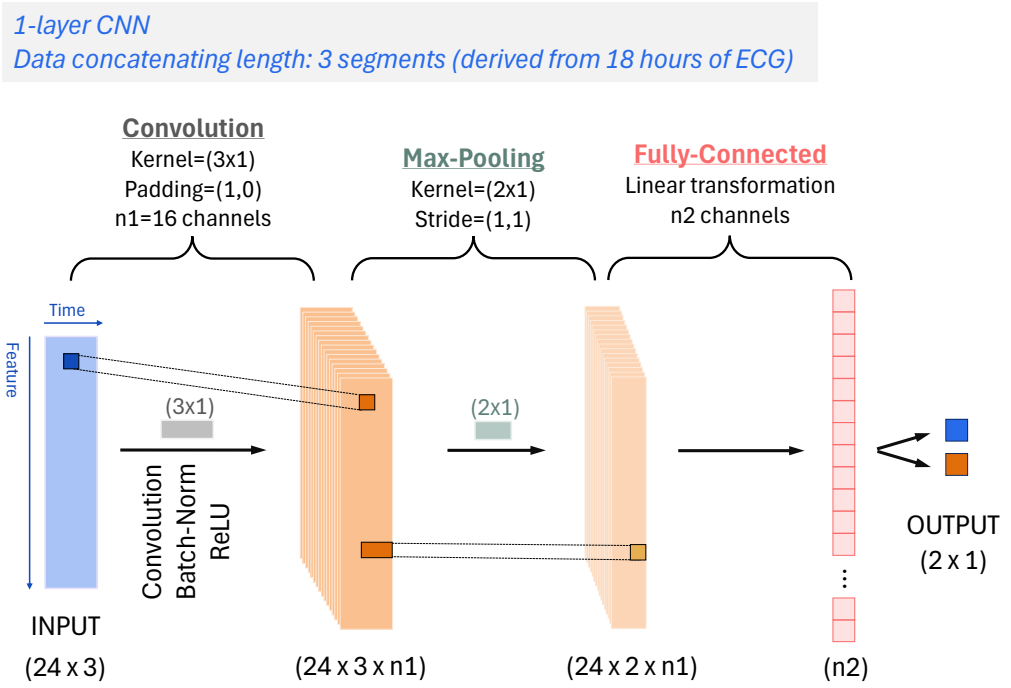

Figure 5.11: Architecture of a shallow CNN with one convolutional layer used in this study.

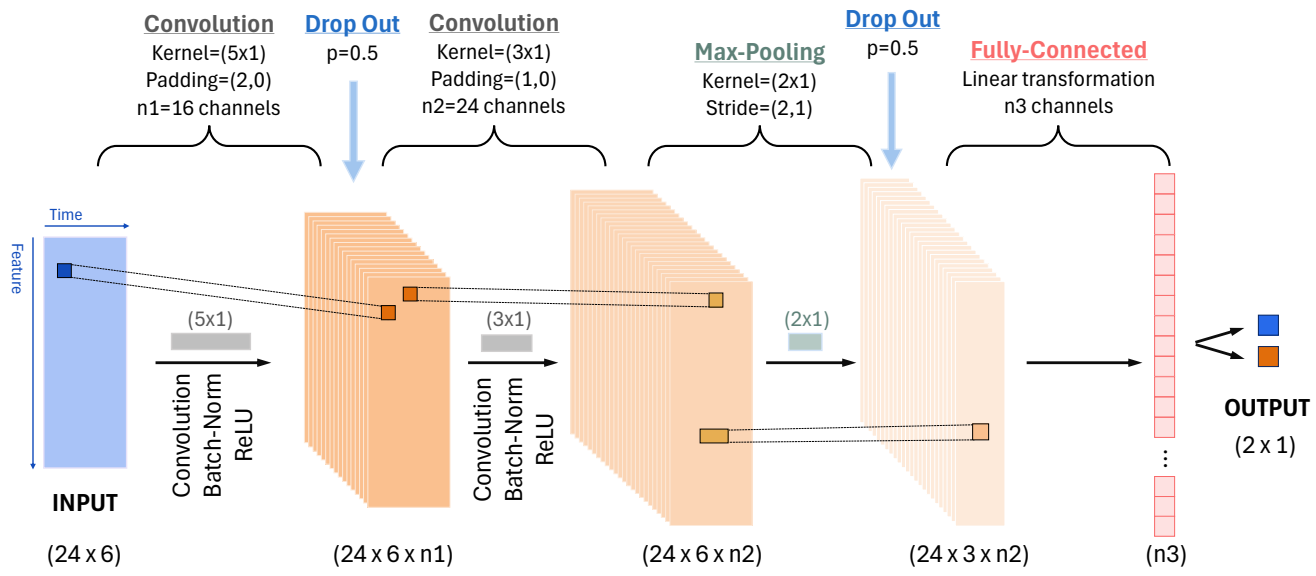

Figure 5.12: Architecture of a shallow CNN with two convolutional layers used in this study.

In the shallow CNN with 1 convolution layer, we used the tensor-version of $ConcatX(3)$ as input, which includes information derived from a total of 18 hours of raw ECG given that the resolution of our data segments was 6 hours. As displayed in Figure 5.11, the inputted data is first processed by a combo of convolution, batch-normalization [48] and non-linear transformation by an activation function (ReLU), then a max-pooling is performed, followed by a linear transformation via a fully-connected layer. The output is the probabilities of two classes: "1" for sepsis and "0" for non-sepsis. And the 2-layer shallow CNN, as illustrated in Figure 5.12 takes a "wider" input tensor which was derived from a concatenation of the current data and 5 more preceding segments, representing 36 hours of ECG signals. This greater input makes it possible to perform two convolutional operations to dig more underlying information that maps to the septic or non-septic status. So, two shallow CNN were thus developed (1-layer and 2-layer).

In summary, we developed and compared a number of detectors to facilitate the early detection of LOS in preterm infants based on different ML algorithms and different feature variants, including original feature sets (*HRV+GA+PNA* that were derived in Section 5.3.1) and the HRV variants described above. Using the original feature sets, we trained **instantaneous** detectors with NB, LR, RF, XGBoost and MLP algorithms. Moreover, we further developed **both**, RF and MLP models with **four HRV variants**: $\Delta X(1)$, $\Delta X(2)$ and $\Delta refX$ and $ConcatX(3)$ to assess the impact of incorporating historical information on sepsis detection. And two shallow CNN were trained with $ConcatX(3)$ and $ConcatX(6)$ variants. Besides, it should also be noted that the manipulations (variants) on the feature sets may introduce a certain loss regarding available sample numbers,

just as these strike-out samples demonstrated in Figure 5.10. Thus, from a fair perspective, models trained by different variants cannot be directly compared because their training data amounts would be different. All algorithms were implemented in Python with *scikit-learn* [49] and *PyTorch* [50] libraries.

5.4.3 Model evaluation metrics

Several metrics were used to measure the ability to detect sepsis of the proposed classifiers. First is Area under the receiver operating characteristic curve (AUROC), which is a comprehensive metric that quantifies the ability of a classifier to distinguish between positive and negative samples at various decision thresholds. It ranges from 0 to 1, where 0.5 approximates a random guess and a higher AUROC indicates better discrimination performance. Area under the precision-recall curve (AUPRC) was also utilized to evaluate the performance in settings with imbalanced class distributions. It measures the trade-off between precision and recall offering insights into the classifier's ability to identify positive instances accurately while minimizing false positives. Note that the baseline of AUPRC depends on the actual proportion of positive instances in the dataset, which is often low in medical scenarios (rare events). Sensitivity (S_e, recall, TPR) and Specificity (S_p, TNR) were also calculated. Besides, Balanced accuracy (BAcc) provides an overall performance metric by considering both sensitivity and specificity, and it calculates the average of both, offering a balanced view of the classification across both positive and negative classes. More information about the metrics calculation is given in Section 2.4.3.

5.4.4 Sensitivity analysis

Sensitivity analysis was performed to study the contributions and interactions of extracted HRV features and clinical information to sepsis detection. The Morris screening method [51], as described in Section 2.5.2, was adopted to analyze single classifiers (detectors).

5.5 Results

We present the results concerning three major aspects: *1)* the performance of various models and their variants that were developed for sepsis detection, *2)* sensitivity analysis of selected best-performing models, and *3)* the behavior and trends of extracted HRV parameters before and during the onset of sepsis (in a subset of the population with sufficient available data).

5.5.1 Sepsis detection performance of the ML models

We evaluated the model performance of detecting the presence of Late-onset sepsis for all the Machine learning algorithms combined with different variants of the feature set, as elaborated in Section 5.4.2. All the reported results were obtained by the "ultimate" Monte Carlo cross-validation

(10 folds) when fixing the hyper-parameters as the optimal sets searched during the "preliminary" process (refer to Section 5.4.1 for the strategy description).

Firstly, the results of time-independent detectors based on five ML algorithms are presented in Table 5.1. In these **instantaneous** detectors, decisions were made solely based on the current segments without considering any past information. In terms of AUROC, all models except Naïve Bayes achieved performances above 0.70, among which RF and MLP have an AUROC of approximately 0.73±0.06. The baseline of AUPRC was 0.123±0.024 in the test data, which is equivalent to the proportion of positive samples (labeled as "1"). The RF and MLP, again, obtained the highest AUPRC of about 0.352±0.099 and 0.340±0.128, respectively. Regarding sensitivity, specificity and their average values, (i.e., Balanced accuracy), MLP and RF performed relatively well.

The Receiver operating characteristic curve (ROC) and Precision-recall curve (PRC) of instantaneous RF and MLP models were plotted in Figure 5.13 and Figure 5.14, respectively. For RF models, their AUROC values ranged from 0.65 to 0.85 with a mean AUROC of 0.727±0.060 across all models as shown in Figure 5.13a. Model #3 achieved the highest AUROC of 0.85. The red dots indicate threshold points along the overall ROC with corresponding False positive rate (FPR) and False negative rate (FNR) values. As the FPR increases, the FNR decreases, reflecting the trade-off between correctly identifying sepsis cases and avoiding false positives. For example, at a low FPR of 0.10, the FNR is 0.62, while when tolerating a higher FPR of 0.30, the FNR drops to 0.36, capturing more true positive cases at the cost of more false positives. From a clinical perspective, a threshold with an FPR of 0.40 and an FNR of 0.27 offers a balance between detecting true positives and managing false positives, but for critical cases, a lower FNR may be more important. Regarding the PRC shown in Figure 5.13b, these curves highlight the trade-off between precision (true positives among predicted positives) and recall (true positives among all actual positives), which is especially important for rare event detection or imbalanced datasets like sepsis detection. The average precision (AP) values of the 10 realizations range from 0.19 to 0.49, with a mean performance of 0.352±0.099 way better than the chance level. These results show that while the random forest models can detect sepsis effectively, there is significant variation in how well they balance precision and recall. For example, at higher recall levels (above 0.6), many models experience a drop in precision, meaning more false positives are introduced.

Shown in Figure 5.14 are the ROC and PRC for MLP models. In general, MLP models achieved similar average performance to RF models. When comparing the key thresholds, there are subtle differences. At lower FPR values (e.g., 0.10 to 0.30), both models exhibit similar TPR trends, but as the FPR increases beyond 0.30, RF models tend to maintain slightly better TPR compared to MLP models, as indicated by slightly lower FNR values at most threshold points. At higher FPR values (e.g., 0.50 to 0.80), the two models converge in performance, showing very similar FNR values.

Table 5.1: Performance of proposed time-independent (**instantaneous**) LOS detectors.

Model	Hyper-Parameters	AUROC	AUPRC (Baseline: 0.123±0.024)	Sensitivity	Specificity	BAcc
NB	/	0.651±0.050	0.210±0.066	0.247±0.121	0.876±0.051	0.644±0.018
LR	{C=2.0, penalty='l2'}	0.720±0.058	0.328±0.124	0.663±0.082	0.669±0.039	0.666±0.038
RF	{nEstim=180, criter='gini', maxdepth=5, maxfea=4, minleaf=6, minsplit=10}	0.727±0.060	0.352±0.099	0.637±0.079	0.692±0.053	0.664±0.043
XGBoost	{nEstim=150, maxdepth=3, lr=0.02, subsamp=0.5, colsamp=0.5, alpha=0.1, lambda=2.0}	0.704±0.067	0.328±0.106	0.611±0.099	0.677±0.062	0.644±0.049
MLP	{hidden=(25), active='tanh', lr='constant', alpha=3.16, solver='adam'}	0.725±0.063	0.340±0.128	0.674±0.084	0.660±0.037	0.667±0.045

The best two results for each metric are marked in light blue and dark blue, respectively.

Next, a set of Δ variants of models based on the best-performing algorithms (RF and MLP) and simple LR were further trained and assessed. The results are shown in Table 5.2. As mentioned in Section 5.4.2, three variants on the HRV values were compared: $\Delta X(1)$, $\Delta X(2)$ and $\Delta refX(r)$. Generally, the $\Delta X(1)$ and $\Delta X(2)$ variants appeared with no added values, or even lower performance, for sepsis detection when compared to their counterparts with the original feature set. The $\Delta refX(r)$, on the other hand, contained the information from a rather longer time ago by utilizing the relative HRV values by subtracting the first non-sepsis segment from the current segment. Roughly, the $\Delta refX(r)$ variants gained better specificity (TNR) at the cost of long-term time dependency and low AUROC, AUPRC and TPR. Across different Δ variants shown in this results table, RF achieved the overall best performance in most metrics.

The remaining set of models was trained with the $ConcatX(n)$ variants, simply stacking multiple preceding data segments and directly using them as the input. Shown in Table 5.3 are the performances of LR, RF and MLP, which were developed with features extracted from three preceding 6-hour segments ($ConcatX(3)$), while the two shallow CNN were developed using $ConcatX(3)$ and $ConcatX(6)$ features. Generally, it seems that with more historical information fed to the models, greater overall performance could be observed. Compared to the results presented in Table 5.1 and Table 5.2, LR, RF and MLP all got satisfactory results, with AUROC of about 0.72 and AUPRC (baseline: 0.12) of 0.32, 0.34 and 0.35, respectively. Moreover, the proposed two shallow CNNs achieved the highest performance: the 1-layer CNN with $ConcatX(3)$ as the input had an AUROC of 0.737±0.027 and an AUPRC of 0.383±0.061 (baseline: 0.130±0.023); and the 2-layer CNN with $ConcatX(6)$ as the input obtained an AUROC of 0.749±0.045 and an AUPRC of 0.378±0.109 (baseline: 0.129±0.015) but with greater variance between cross-validation. In terms of the rest metrics, 1-layer CNN with $ConcatX(3)$ variant had the best sensitivity (recall, True positive rate) of 0.676±0.072, and the 2-layer CNN with $ConcatX(6)$ variant had the best specificity (True negative rate) of 0.815±0.065. Overall, the shallow CNNs obtained the best Balanced accuracy (BAcc) of 0.667±0.021 and 0.678±0.049, respectively.

Figure 5.15 and Figure 5.16 show the receiver operating characteristic curves and precision-recall curves of the 1-layer CNN and 2-layer CNN model, respectively. Apart from achieving one of the best overall performances, the 1-layer CNN models also demonstrate the most robust performance in terms of both AUROC and AUPRC. This is evidenced by the smaller standard deviations across 10 different realizations, as indicated by the colored lines in Figure 5.15a and Figure 5.15b.

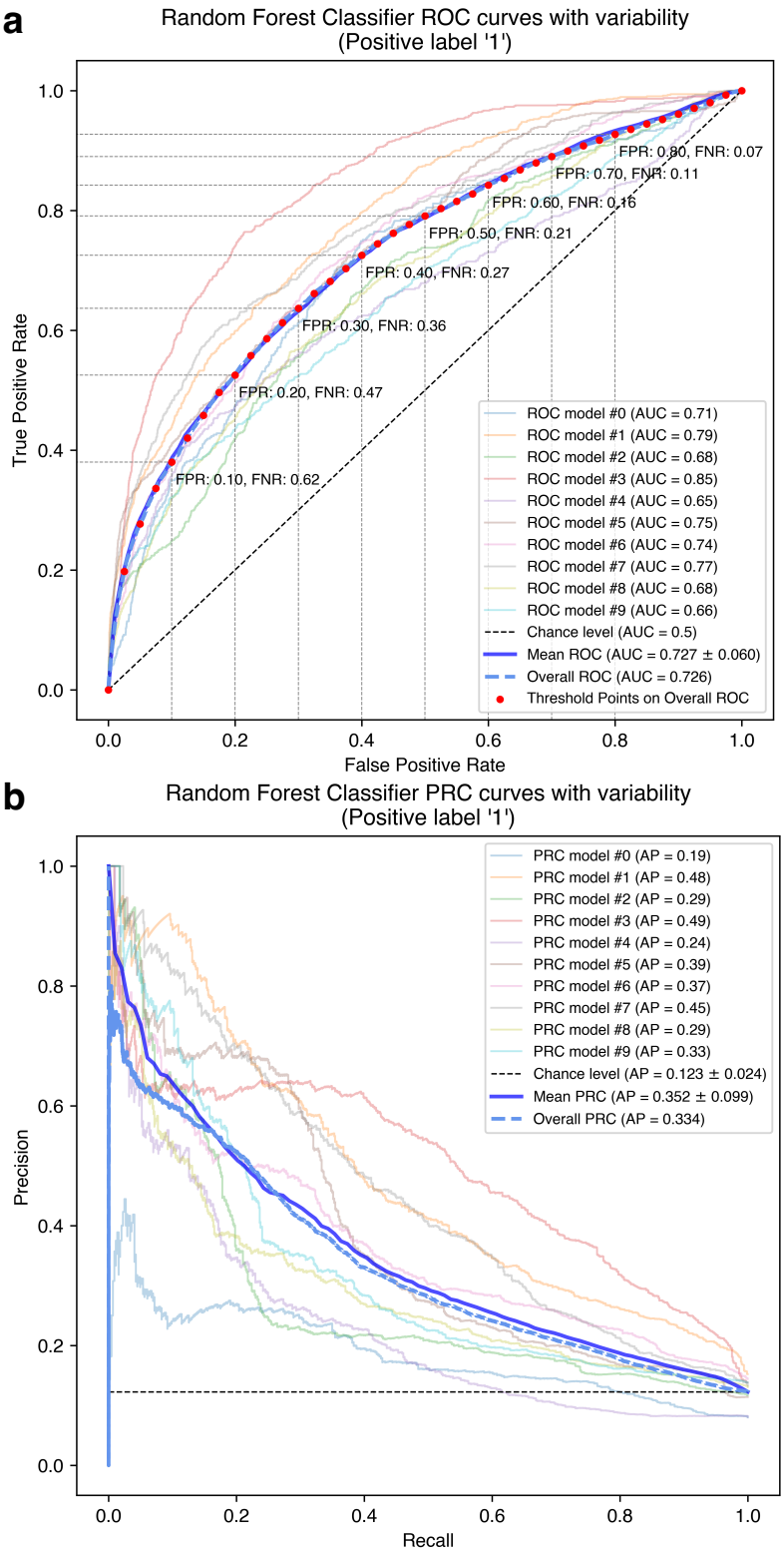

Figure 5.13: Performance of an instantaneous random forest detector in Monte Carlo cross-validation. (**a**) Receiver operating characteristic curves (ROC) in 10 realizations. The thick blue line indicates the mean ROC curve across all models, with an average AUC of 0.727±0.060. The dashed black line shows the performance of a random guess (AUC=0.5). Red dots indicate threshold points along the overall ROC curve, with corresponding FPR (false positive rate) and FNR (false negative rate) values labeled for reference. (**b**) Precision-Recall curves (PRC) in 10 realizations. The thick blue line indicates the mean PRC curve across all models, with an average AUC of 0.352±0.099. The dashed black line shows the chance level (AP=0.123±0.024).

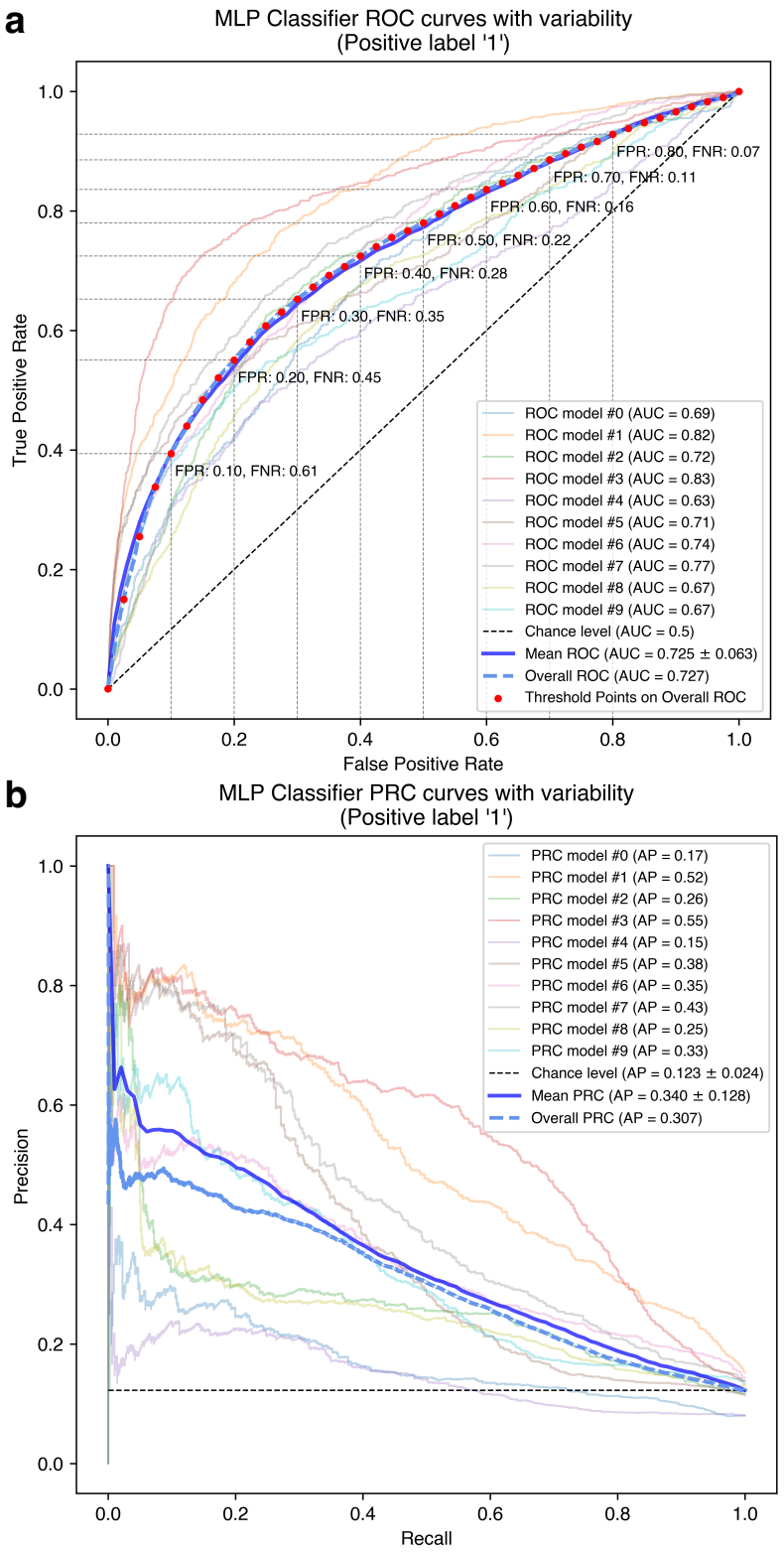

Figure 5.14: Performance of an instantaneous multilayer perceptron detector in Monte Carlo cross-validation. (**a**) Receiver operating characteristic curves (ROC) in 10 realizations. The thick blue line indicates the mean ROC curve across all models, with an average AUC of 0.725±0.063. The dashed black line shows the performance of a random guess (AUC=0.5). Red dots indicate threshold points along the overall ROC curve, with corresponding FPR (false positive rate) and FNR (false negative rate) values labeled for reference. (**b**) Precision-Recall curves (PRC) in 10 realizations. The thick blue line indicates the mean PRC curve across all models, with an average AUC of 0.340±0.128. The dashed black line shows the chance level (AP=0.123±0.024).

Table 5.2: Performance of proposed time-dependent (**Δ variants**) LOS detectors.

HRV	Model	Hyper-Parameters	AUROC	AUPRC	Sensitivity	Specificity	BAcc
$\Delta X(1)$	LR	{C=0.001, penalty='l2'}	0.652±0.059	(Baseline: 0.128±0.025) 0.212±0.103	0.667±0.109	0.574±0.069	0.621±0.044
	RF	{nEstim=200, criter='gini', maxdepth=4, maxfea=3, minleaf=8, minsplit=7}	0.678±0.024	0.233±0.055	0.669±0.086	0.567±0.064	0.618±0.025
	MLP	{hidden=(50,50,50), active='relu', lr='constant', alpha=0.001, solver='sgd'}	0.675±0.040	0.250±0.073	0.635±0.074	0.617±0.046	0.626±0.034
$\Delta X(2)$	LR	{C=0.01, penalty='l2'}	0.634±0.050	(Baseline: 0.119±0.016) 0.177±0.044	0.594±0.138	0.589±0.059	0.591±0.048
	RF	{nEstim=100, criter='gini', maxdepth=4, maxfea=3, minleaf=5, minsplit=5}	0.690±0.044	0.244±0.047	0.689±0.073	0.576±0.040	0.632±0.032
	MLP	{hidden=(75,), active='relu', lr='adaptive', alpha=0.1, solver='sgd'}	0.678±0.042	0.240±0.048	0.627±0.078	0.645±0.037	0.636±0.029
$\Delta refX(r)$	LR	{C=0.01, penalty='l2'}	0.670±0.025	(Baseline: 0.099±0.025) 0.183±0.051	0.589±0.087	0.672±0.060	0.631±0.027
	RF	{nEstim=160, criter='gini', maxdepth=6, maxfea=5, minleaf=7, minsplit=10}	0.703±0.018	0.215±0.037	0.519±0.072	0.740±0.041	0.630±0.019
	MLP	{hidden=(50,), active='relu', lr='constant', alpha=0.01, solver='sgd'}	0.699±0.035	0.241±0.070	0.562±0.085	0.734±0.049	0.648±0.029

The best results for each variant and metric are marked in light blue.

Table 5.3: Performance of proposed time-dependent (***Concat* variants**) LOS detectors.

Model	HRV	Hyper-Parameters	AUROC	AUPRC (Baseline)	Sensitivity	Specificity	BAcc
LR	$ConcatX(3)$	{C=0.1, penalty='l2'}	0.720±0.041	0.317±0.044 (0.118±0.016)	0.591±0.067	0.741±0.030	0.666±0.033
RF	$ConcatX(3)$	{nEstim=120, criter='gini', maxdepth=7, maxfea=5, minleaf=6, minsplit=10}	0.717±0.051	0.340±0.062 (0.118±0.016)	0.545±0.075	0.769±0.035	0.657±0.039
MLP	$ConcatX(3)$	{hidden=(75,), active='relu', lr='constant', alpha=0.01, solver='sgd'}	0.722±0.043	0.345±0.055 (0.118±0.016)	0.587±0.080	0.740±0.027	0.663±0.039
CNN	$ConcatX(3)$	{1-layer: conv+BN+ReLU+Pool+FC}	0.737±0.027	0.383±0.061 (0.130±0.023)	0.676±0.072	0.657±0.078	0.667±0.021
	$ConcatX(6)$	{2-layer: conv1+BN1+ReLU+DO1 +conv2+BN2+ReLU+Pool+DO2+FC}	0.749±0.045	0.378±0.109 (0.129±0.015)	0.541±0.119	0.815±0.065	0.678±0.049

The best two results for each metric are marked in light blue and dark blue, respectively.

5.5.2 Sensitivity analysis

The Morris sensitivity analysis of the developed sepsis detection models provides insights into the relative importance and their interactions among various features.

Shown in Figure 5.17 is the sensitivity analysis of an RF instantaneous detector (#3 realization of 10 times of Monte Carlo validation). The left panel illustrates the feature importance (solid bars) stacked with feature interactions (hatched bars) in the model, while the right panel of the figure further displays each feature on the $\mu^* - \sigma$ plane. Different features were presented in different color palettes: redish for ages, purplish for time-domain HRV, yellowish for frequency-domain HRV and greenish for non-linear HRV.

Seeing from the bar plot, the most influential features in the RF model are Postnatal age (PNA), Gestational age (GA), mean, pDec (percentage of deceleration), median of RR intervals and DC (decelerate capacity), reflecting their strongest ability to influence predictions both independently and through interactions with other features. This can also be observed on the right panel of the figure, in which these features are positioned toward the upper right. The second tier of influential features are skewness, the non-linear HRV parameters including SampleEn (sample entropy), $\alpha 2$ and $\alpha 1$ from detrended fluctuation analysis, and two frequency-domain HRV related to LF (low-frequency power). They exhibit moderate importance with relatively higher interaction effects. Most of the remaining features are clustered in the lower-left corner of the right panel as shown in the scatter plot, indicating their limited effects on the model's overall detection for the onset of LOS.

5.5.3 Changes in HRV around the sepsis onsets

The dynamics of all extracted features within 5 days before and after the onset of confirmed sepsis (denoted by "T0") are visualized in Figure 5.18 and Figure 5.19. These trends were generated from a selected subset of the overall database, consisting of 84 sepsis episodes from 73 patients. To ensure a representative analysis of the transition from pre-sepsis to sepsis, only episodes with sufficient duration covering both the septic states and the preceding control states were included. Overall, the plots show noticeable shifts in several HRV features as the patients transition from the pre-sepsis (in green, denoted by "Label=0") to the sepsis state (in red, denoted by "Label=1"). The wider error bars in the two ends of the plots were due to the lack of data.

In terms of time-domain HRV features shown in Figure 5.18, both the mean and median HRV (first two subplots in Figure 5.18) show relatively stable values in the control period. However, starting from 6 hours before the onsets of sepsis (T0), a drop in their values is observed, followed by an evident elevation 6 hours after sepsis onsets. Similar but more persistent, the skewness (row 2, col 1) presents a drastic decrease from 6 hours before the sepsis onset until 30 hours after the onset time, then it raises gradually to the normal level. AC (row 3, col 2) and DC (row 3, col 2) show

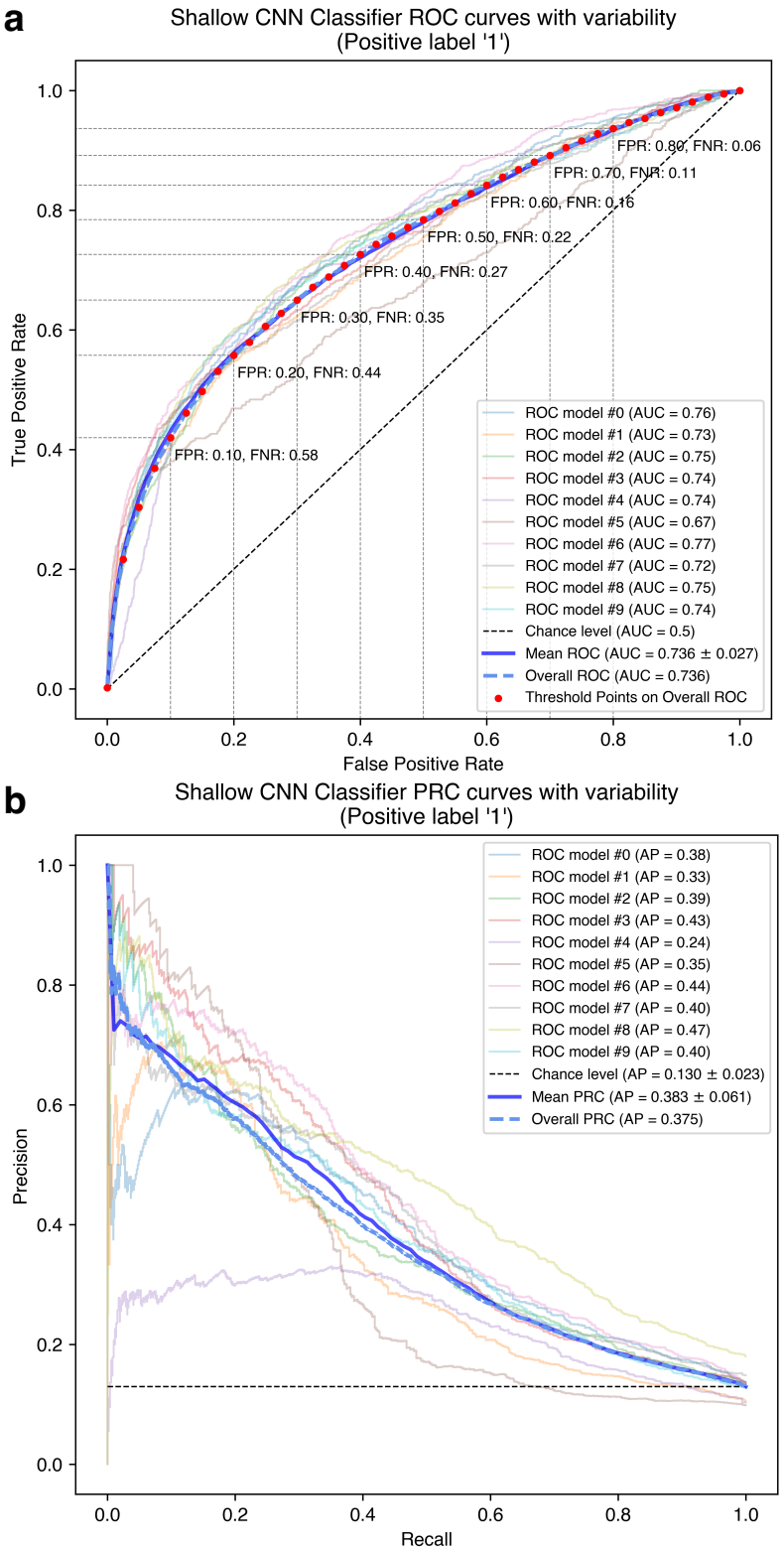

Figure 5.15: Performance of time-dependent 1-layer shallow CNN detector in Monte Carlo cross-validation. (**a**) Receiver operating characteristic curves (ROC) in 10 realizations. The thick blue line indicates the mean ROC curve across all models, with an average AUC of 0.736±0.027. The dashed black line shows the performance of a random guess (AUC=0.5). Red dots indicate threshold points along the overall ROC curve, with corresponding FPR (false positive rate) and FNR (false negative rate) values labeled for reference. (**b**) Precision-Recall curves (PRC) in 10 realizations. The thick blue line indicates the mean PRC curve across all models, with an average AUC of 0.383±0.061. The dashed black line shows the chance level (AP=0.130±0.023).

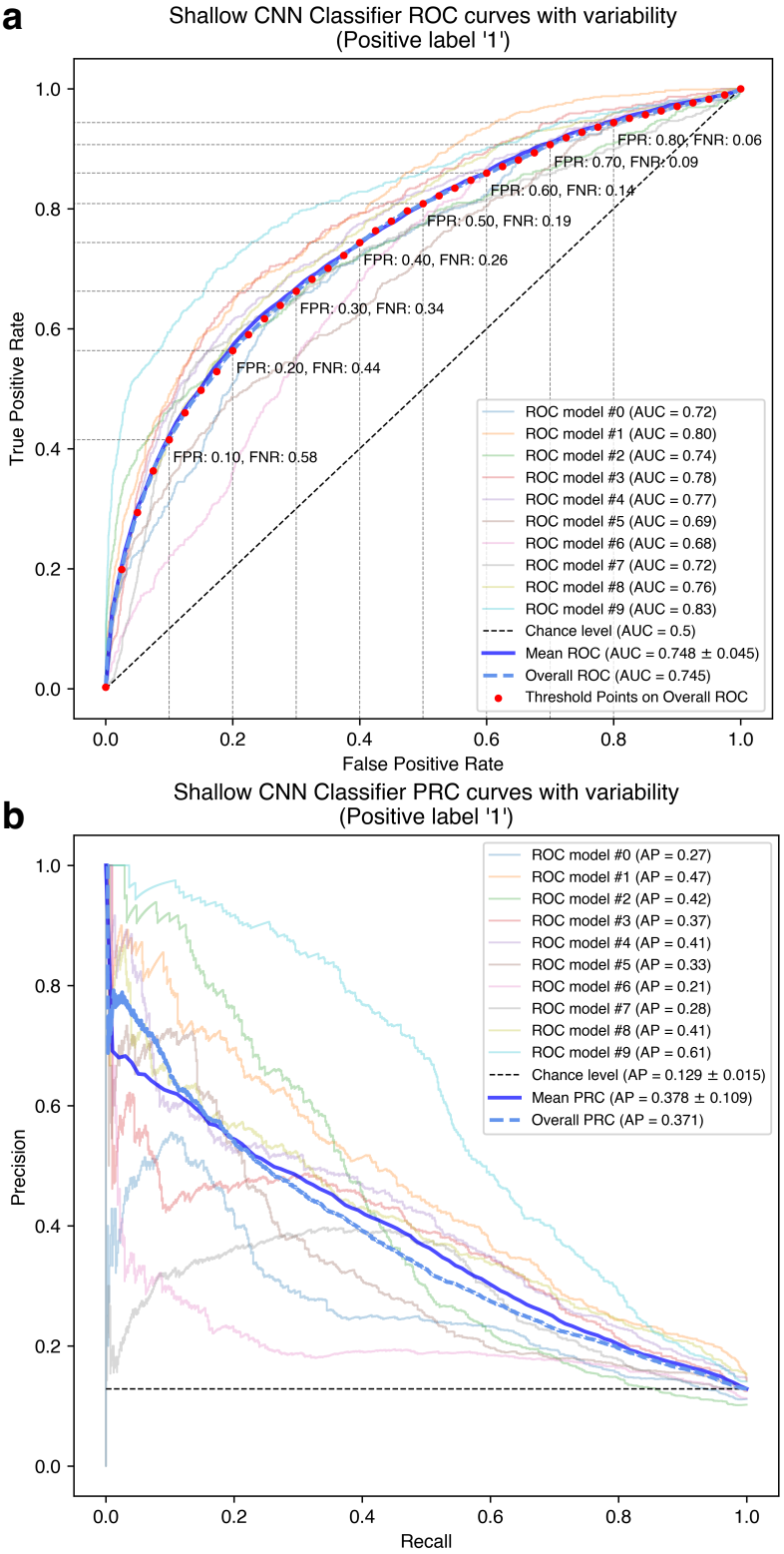

Figure 5.16: Performance of time-dependent 2-layer shallow CNN detector in Monte Carlo cross-validation. (**a**) Receiver operating characteristic curves (ROC) in 10 realizations. The thick blue line indicates the mean ROC curve across all models, with an average AUC of 0.748±0.045. The dashed black line shows the performance of a random guess (AUC=0.5). Red dots indicate threshold points along the overall ROC curve, with corresponding FPR (false positive rate) and FNR (false negative rate) values labeled for reference. (**b**) Precision-Recall curves (PRC) in 10 realizations. The thick blue line indicates the mean PRC curve across all models, with an average AUC of 0.378±0.109. The dashed black line shows the chance level (AP=0.129±0.015).

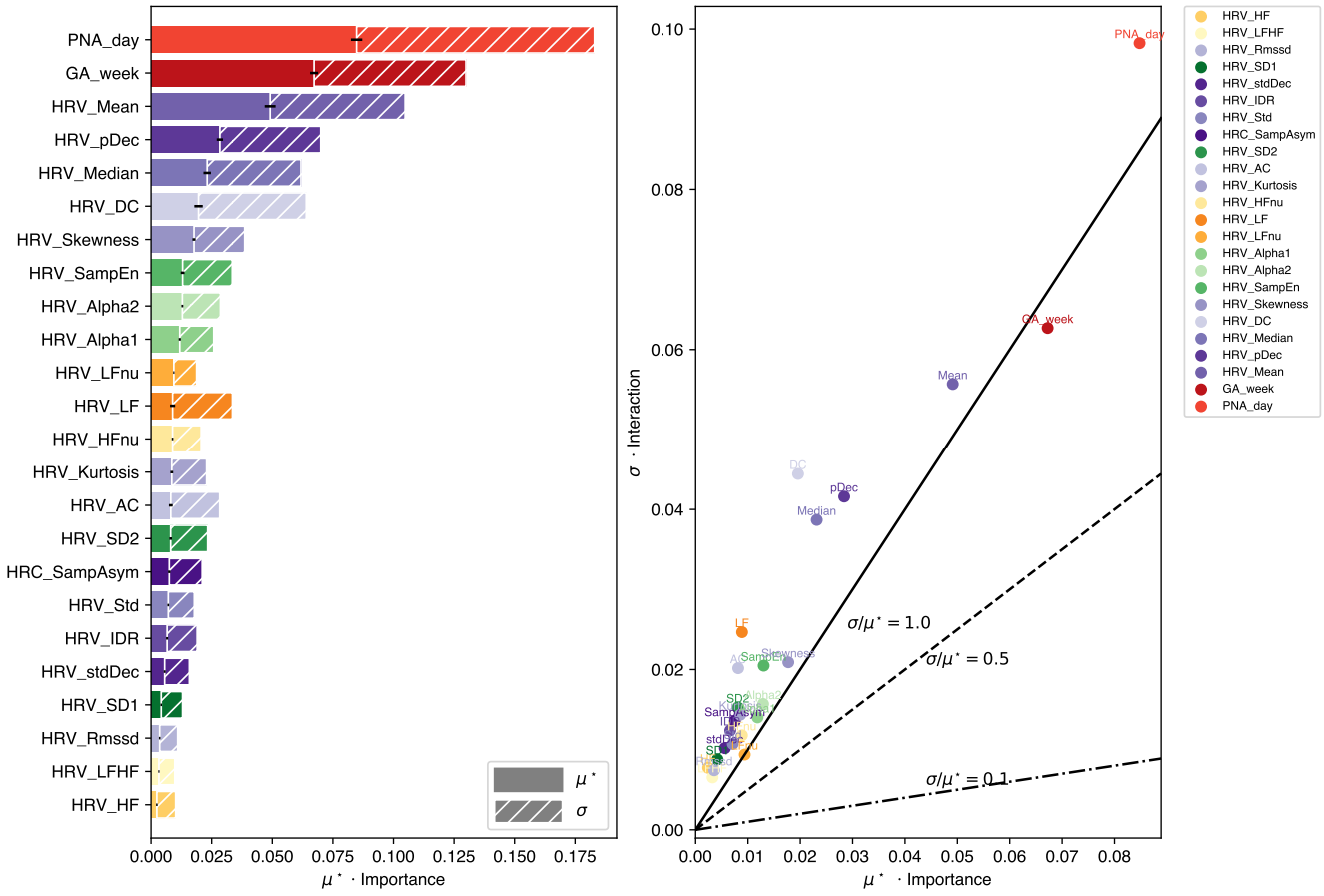

Figure 5.17: Sensitivity analysis of an RF instantaneous detector (#3 realization of Monte Carlo validation). The left panel illustrates the feature importance (μ^*, solid bars) stacked with feature interactions (σ, hatched bars) in the model, while the right panel of the figure further displays each feature in scatter on the $\mu^* - \sigma$ plane. Different features were presented in different color palettes: redish for ages, purplish for time-domain HRV, yellowish for frequency-domain HRV and greenish for non-linear HRV.

similar decreases in advance and it lasts until 12 hours after the onset of sepsis. On the contrary, before the sepsis onsets, the value of the percentage of deceleration (pDec, raw 4, col 1) shows a decreasing trend and bottoms out at around 6 hours before the onsets and then keeps increasing until 30 hours in sepsis states. Another obvious dynamic is observed from sample asymmetry (row 4, col 3), at approximately 24 hours before sepsis, the value starts to rise and peaks at 6 hours prior to the onsets, then it drops steadily until 24 hours of sepsis to a lower level.

The transition trends of frequency-domain and non-linear HRV features are shown in Figure 5.19. Both LF (low-frequency power, row 1, col 1) and normalized LF (row 1, col 2) present a decrease from 6 hours before sepsis onset to around 12-18 hours after. Sample entropy (row 3, col 2) shows an upward trend from the onset of sepsis till 18 hours later. $\alpha 2$ (row 4, col 1) is observed with a significant drop 18 hours preceding sepsis onset and starts to increase until approximately 24 hours after being diagnosed with sepsis.

Moreover, it is known that gestational age plays a crucial role in the context of neonatal care as it directly influences the physiological development of neonates' vital systems and may lead to variable pathological courses. To get closer to the HRV dynamics without mixing with possible interaction with different GA, we further present the HRV transition trends by stratifying the selected population into GA$\leq$28 weeks (54 extremely preterm infants) and GA>28 weeks (19 very preterm infants), as shown in Figure 5.20 and Figure 5.21. Generally, despite the differences in patient numbers of the two age groups, the heart rate variability in the older group (lines in light green and light red) exhibits more dynamic and evident alterations along the observation duration, while the variability measured in the younger group (lines in green and red) is much more subtle.

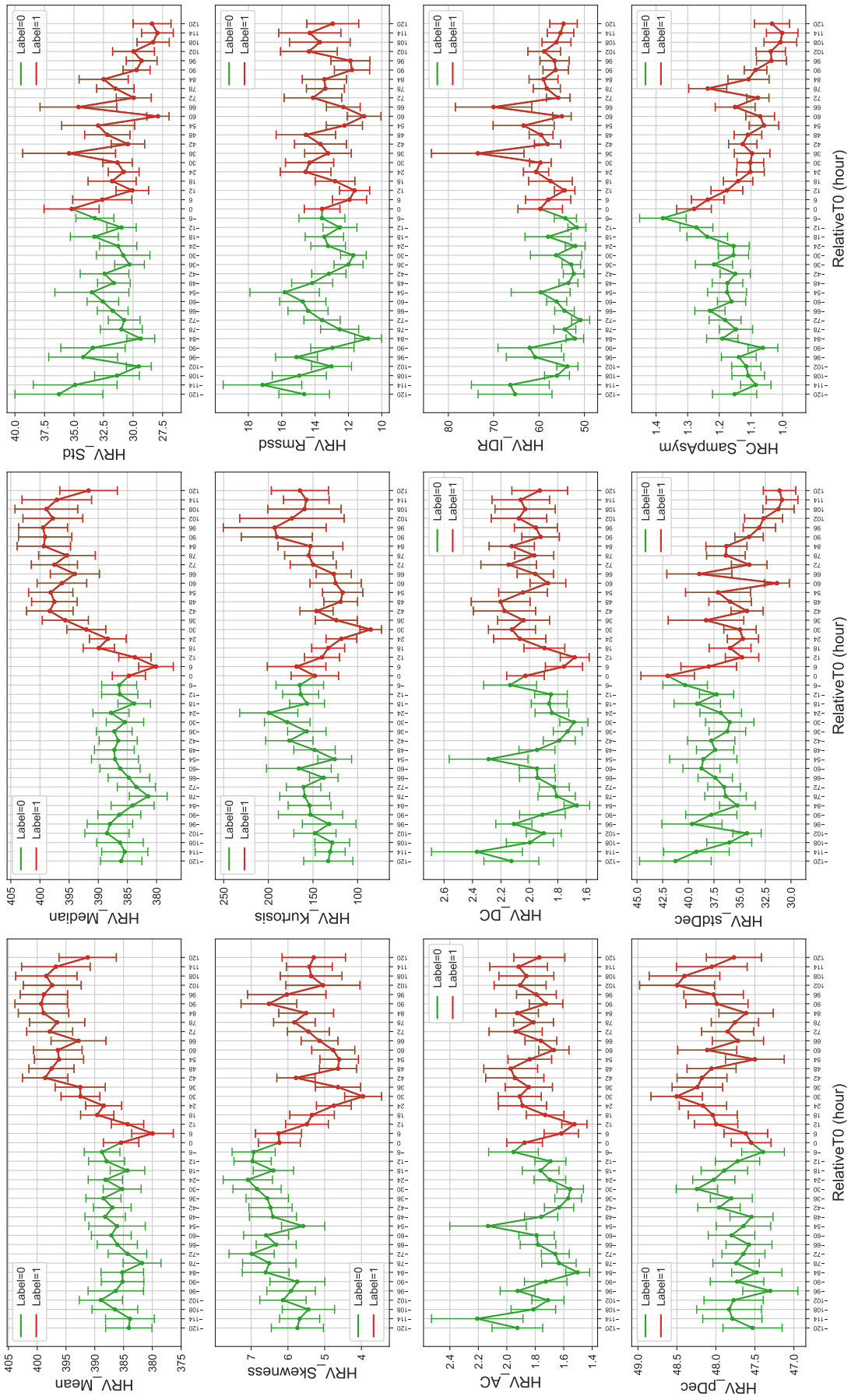

Figure 5.18: Changes in HRV features within 5 days before and after the annotated LOS onsets (1/2). Statistics were calculated from selected events (84 confirmed sepsis episodes from 73 patients) that have sufficient duration in both non-septic (in green, denoted by "Label=0") and septic states (in red, denoted by "Label=1"). The scatters represent the average feature values at the event level. The error bars represent the Standard error of the mean (SEM).

Figure 5.19: Changes in HRV features within 5 days before and after the annotated LOS onsets (2/2). Statistics were calculated from selected events (84 confirmed sepsis episodes from 73 patients) that have sufficient duration in both non-septic (in green, denoted by "Label=0") and septic states (in red, denoted by "Label=1"). The scatters represent the average feature values at the event level. The error bars represent the Standard error of the mean (SEM).

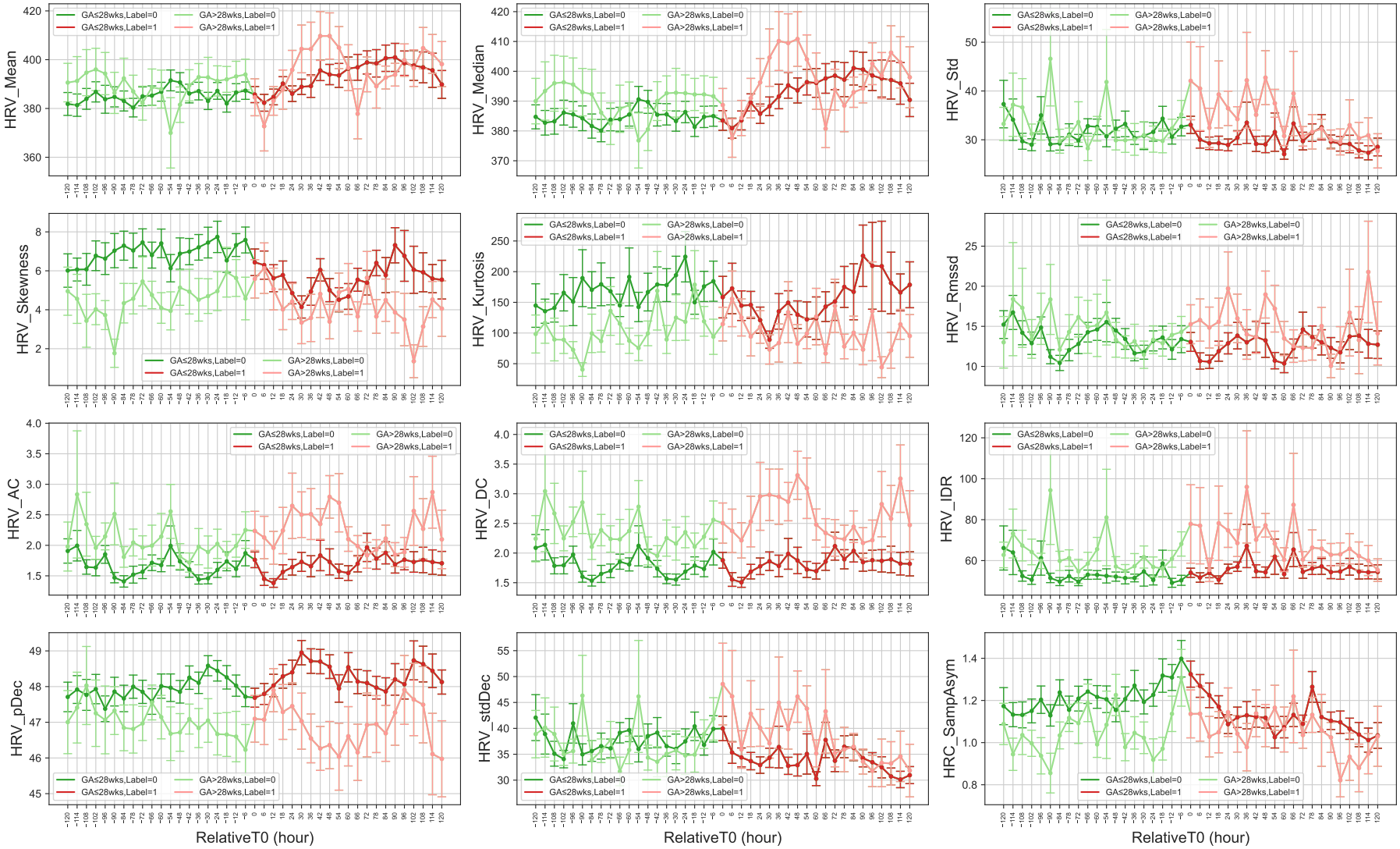

Figure 5.20: Changes in HRV features by GA within 5 days before and after the annotated LOS onsets (1/2). The green (non-sepsis, denoted by "Label=0") and red (sepsis, denoted by "Label=1") lines present the HRV trends in patients with GA≤28 weeks (calculated from 64 confirmed sepsis episodes from 54 patients). The light green (non-sepsis, denoted by "Label=0") and light red (sepsis, denoted by "Label=1") lines present the HRV trends in those with GA>28 weeks (calculated from 20 confirmed sepsis episodes from 19 patients). The scatters represent the average feature values at the event level. The error bars represent the Standard error of the mean (SEM).

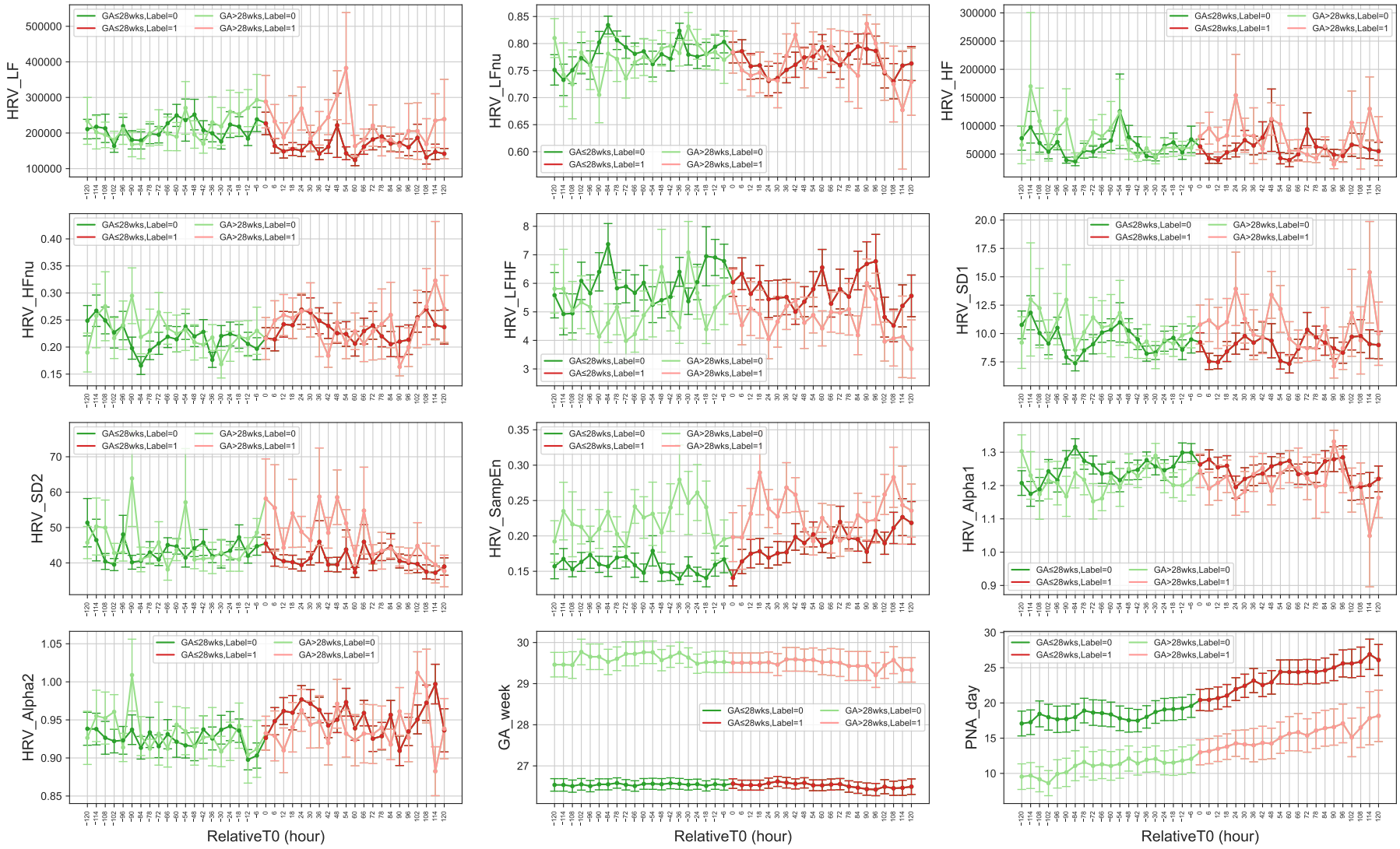

Figure 5.21: Changes in HRV features by GA within 5 days before and after the annotated LOS onsets (2/2). The green (non-sepsis, denoted by "Label=0") and red (sepsis, denoted by "Label=1") lines present the HRV trends in patients with GA≤28 weeks (calculated from 64 confirmed sepsis episodes from 54 patients). The light green (non-sepsis, denoted by "Label=0") and light red (sepsis, denoted by "Label=1") lines present the HRV trends in those with GA>28 weeks (calculated from 20 confirmed sepsis episodes from 19 patients). The scatters represent the average feature values at the event level. The error bars represent the Standard error of the mean (SEM).

5.6 Discussion

In this study, we developed several non-invasive and near real-time ML detectors to discriminate between time intervals when premature infants were in the LOS state versus the control state. Different components of our approach are directly aimed at integrating proposed ML models into clinical care. As in all ML works, these data preparation, database querying, signal processing, feature extraction and constitution of the corpus for training and testing represented a significant part of the work.

By leveraging real-life monitoring ECG signals routinely collected from NICU in multiple clinical centers, we extracted informative HRV parameters and basic clinical information using signal processing techniques. This approach allowed us to capture subtle physiological changes associated with sepsis, facilitating timely and accurate detection in clinical settings.

The raw signal was retrieved from routine NICU monitoring, starting from day 3 of life and continuing throughout the entire hospitalization in the NICU. This led to a longitudinal dataset that captured dynamic physiological changes over time, providing valuable insights into early detection of neonatal LOS.

Through a well-crafted labeling strategy based on a multi-expert clinical classification of neonatal sepsis, we annotated the time series intervals every 6 hours as "1" (sepsis), "0" (non-sepsis) or "-1" (uncertain, were excluded from analysis). Various ML classifiers with different feature set variants were developed and evaluated under a Monte Carlo cross-validation scheme. The best-performing instantaneous model was the RF trained with 22 extracted Heart rate variability (HRV) features and two age indicators, Gestational age (GA) and Postnatal age (PNA). Furthermore, the shallow CNN demonstrated improved performance when detecting sepsis by utilizing data from the previous 18 hours (three consecutive 6-hour segments), highlighting that incorporating temporal information enhances detection accuracy, though at the cost of increased time dependency, making the model less suitable for real-time detection.

The overall observed performances for detecting sepsis in this study are not optimal yet and appear to be lower than findings in other studies [29, 30, 32, 34], but show the potential of using routine monitoring signals to detect LOS in a near real-time manner. In fact, it is important to note that comparisons between studies are complicated, and we claim that it is impossible to compare, due to discrepancies in cohort selection, sepsis definitions, data processing, labeling strategies, and model performance evaluation criteria. These differences lead to wide variability and huge heterogeneity in the performances in the literature. Nevertheless, each study, including ours, provides valuable insights from different perspectives, contributing to advancing early detection methods for LOS.

5.6.1 Different definitions of neonatal sepsis and its onset moment

Unlike pediatric and adult sepsis, there is still no commonly agreed definition for neonatal sepsis [52–55]. While widely recognized criteria like the Vermont Oxford Network criteria [56], NEO-KISS [37] and European Medicines Agency definition [57] are commonly applied, the absence of a standardized definition poses challenges. This lack of consensus has led to variations in how the onset of LOS is identified and managed across different studies.

Some studies have defined the onset of LOS as the time when the administration of antibiotics begins [34], while others use the time a blood culture is drawn as registered in the patient records [58, 59]. Although these are convenient and straightforward markers reflecting the clinician's suspicion of infection, they introduce subjectivity as the timing of antibiotic administration or blood draw for blood cultures relies a lot on clinical judgment, the perceived urgency of the patient's condition, logistical factors like the availability of resources and so on.

The CRASH moment (Cultures, Resuscitation, and Antibiotics Started Here), introduced by Griffin and Moorman [11], provides a more structured approach to defining sepsis onset. CRASH captures the moment when clinicians initiate a set of critical interventions: collecting cultures to identify pathogens, starting resuscitation to stabilize the neonate, and administering antibiotics as an empirical treatment even before culture results are available. Adopted in subsequent studies on neonatal sepsis [32, 33], the CRASH serves as a practical clinical anchor for defining the onset of LOS in preterm infants. However, despite its utility, the reproducibility and the synchronization of these interventions (and, thus, of the CRASH moment) is very limited. For instance, antibiotics may be administered before or after cultures are taken; resuscitation may be initiated based on the infant's condition; different sites may have very different strategies for blood sampling, with different mean times between the blood sampling and the biological results and these mean times may even vary within the same center because of very different causes... All these factors introduce inaccuracies when using CRASH as a precise onset marker.

In our study, we sought to use a consensus-based, formalized approach to define the onset of sepsis to alleviate the above limitations. All suspicious sepsis events were firstly classified retrospectively through a three-expert consensus procedure following the NEO-KISS protocol [37], ensuring systematic classification. We then defined the onset of a sepsis episode as the earlier one between the initiation time of antibiotic treatment and the time of C-reactive protein$>$5 mg/L or positive blood culture (often documented as the time of performing blood sampling). This method takes the very first response of experienced and professional clinicians at the sepsis onset moments. We consider that this method might provide more clinically meaningful results since it minimizes the sources of variance in the annotations, as described above. However, this definition is certainly more challenging as a target for an Machine learning method.

Concerning the pseudo-labeling strategies in this study, we designed an *in-place* logic rule to assign labels to each sample (a 6-hour data segment) based on the event classification from

multi-expert consensus, the simultaneous presence of antibiotic therapy and CRP >5 mg/L of the current data segment. At this stage, we would like to learn detectors that can infer whether or not the patient will develop LOS at the current moment in time, thus, the models were trained on data without shifting the labels earlier. Importantly, to minimize the risk of training ML models on false-positive or false-negative LOS episodes, we applied a labeling strategy that sets strict criteria for septic states and control states, any data that was not conclusively sure to be septic or non-septic were additionally labeled as "-1" and excluded from the analysis. By excluding uncertain samples, the developed models avoid ambiguous situations that might confuse the learning process, leading to improved precision and reduced noise in the training data theoretically.

Generally, in studies targeting the early detection of sepsis, the common labeling manner is "ahead labeling", meaning that a certain number of samples leading up to the identified onsets (t_0) are labeled as positive (sepsis) in advance. For instance, the PhysioNet/Computing in Cardiology Challenge 2019 called for the challenge of early prediction of sepsis from clinical data [60, 61], they labeled the hourly data for septic patients from 6 hours before identified sepsis onsets as "1". Some other studies advanced the positive labels before sepsis onsets ranging from 3 to 48 hours [34, 59], or considered the positive window length (number of hours before onset in LOS patients as positive samples) as a model hyper-parameter and optimized it with cross-validation [30, 32]. This strategy attempts to capture subtle, early physiological changes that occur before clinical signs of sepsis are evident enough to arouse clinical suspicion and intervention. It aligns with the goal of early detection, potentially improving the model's sensitivity. Yet, it should be minded that ahead-labeling may introduce bias into the model, making it more likely to label earlier samples as sepsis-positive even if they do not have enough evidence to suggest an imminent infection. This might lower model specificity, possibly resulting in antibiotic abuse and increased strain on healthcare resources.

5.6.2 Different types of data sources and predictors

In current literature, the proposed algorithms for LOS detection were based on a variety of data sources, including demographic data, vital signs (e.g., heart rate, blood pressure), clinical variables, laboratory tests, Electronic health records (EHR), signal data (e.g., ECG), etc., leading to different types of predictors (features) that were used for developing the LOS early detection models.

Many research on LOS detection through ML algorithms includes the use of data available in EHR. Some researchers have used all recorded variables in EHR, including invasive laboratory results, medications and so on, to train diagnostic models [62–64]. While EHR provides a wealth of information, they present several challenges such as data sparsity, a high ratio of missing values, high dimensionality and inherent data biases. To avoid these issues, some studies have narrowed the feature sets to include only routinely monitored vital signs directly from bedside monitors, or

EHR data with relatively high-resolution recordings. Using high-resolution vital signs (e.g., 0.5 Hz or 1 Hz) such as heart rate, respiratory rate and oxygen saturation has shown promise in early detecting LOS [32, 58, 65]. However, these intermediate variables are often recorded as discrete snapshots and may overlook subtle physiological changes that occur over time.

A more comprehensive approach involves using raw physiological waveform data that are continuously monitored throughout the hospitalization in NICU settings. This data allows for a more detailed and continuous analysis of the patient's disease progression status, capturing subtle changes that might be missed with traditional EHR-based data. In this study, we took advantage of continuous ECG data, which is routinely collected in modern, increasingly digitalized NICU environment. However, real-world data are often prone to noise and inconsistencies, underscoring the importance of robust data pre-processing steps such as denoising and validating data quality. In our study, we applied a comprehensive data processing pipeline to ensure the integrity and reliability of the ECG data and extracted HRV features before model development, as this step is crucial to achieving accurate sepsis detection in clinical practice.

Additionally, longitudinal analysis of other physiological data streams, such as respiratory signals, may offer further diagnostic value for predicting LOS. Recent studies have explored multimodal sepsis prediction models based on various signal "channels". The integration of ECG for cardiac activity, chest impedance waveforms for respiratory activity and motion data for the estimate of infant movement, together with basic clinical information, appeared to be a promising tool to facilitate early detection of neonatal LOS. Joshi et al. [33] analyzed 49 infants with culture-proven LOS by extracting features in 3-hour time intervals including HRV, respiration and ECG-derived estimates of movement, predicting sepsis using a Naïve Bayesian model. Their results demonstrated an increased propensity toward pathological heart rate decelerations, increased respiratory instability and a decrease in spontaneous infant activity (lethargy) in the hours leading up to the clinical suspicion of sepsis. Similarly, Cabrera-Quiros et al. [29] used a combination of HRV, respiration and body motion features extracted every 1-hour interval in both sepsis and age-matched control infants to develop three ML models (Logistic Regression, Naïve Bayes and Nearst mean classifier), achieving a mean accuracy of 0.79 ± 0.12 and a precision rate of 0.82 ± 0.18 at three hours before sepsis onset.

Further, Peng et al. [30] enriched the variety of informative features to 60 HRV, 35 respiration features, and 35 motion features, extracted every hour for the 24 hours before the onset of sepsis. Using multiple ML algorithms, they reported that the XGBoost classifier trained with all extracted features (together with gestational age and birth weight) achieved the best performance, with an AUROC of 0.88 ± 0.09 during the six hours preceding LOS onset. Subsequently, Yang et al. [32] expanded on the work in [30] by using second-by-second vital signs to develop 7 AI models for LOS risk prediction. With the same idea of HRV, they calculated a series of features based on heart rate, respiratory rate and oxygen saturation. Their best-performing model, XGBoost classifier, reached an AUROC of 0.875 ± 0.072 in 7-fold cross-validation for predicting LOS during the 6 hours be-

fore sepsis moments. An additional contribution of this study was the proposition of a multi-level alarm strategy, which may facilitate patient stratification management in clinical use and minimize alarm fatigue, making the model more applicable for clinical use.

In terms of the generalization of HRV in clinical practice, although ECG monitoring is widely available, HRV analysis has yet to be integrated into routine practice, largely due to proprietary software, hardware requirements and significant upfront costs. Additionally, healthcare professionals need proper training to interpret HRV data and understand its implications for sepsis diagnosis and prognosis, which poses further barriers to adoption [10]. In the context of neonates, particularly preterm infants, additional complexities arise when analyzing HRV signals. Artifacts in the data must be managed, requiring robust post-processing capabilities for accurate interpretation. Moreover, abnormal HRV patterns can result from factors beyond sepsis, such as gestational age or other underlying medical conditions [66], and these influences must be carefully considered during analysis to avoid misinterpretation [10].

5.6.3 Clinical utility of sepsis early detection models

The clinical utility goes beyond raw performance, and strong results do not always imply real-world usability. For the clinicians to correctly interpret the model performance, we reported several evaluation metrics. In the ML literature, it is widely recognized that the commonly used AUROC metric, despite its comprehensiveness, tends to give an overly optimistic assessment of model performance in class-imbalanced datasets. This is because AUROC is disproportionately influenced by a large number of true negative predictions [67]. However, most clinical studies based on ML techniques continue to report AUROC as the primary metric for evaluating their overall discriminatory performance on snapshots of the full time-series dataset [59]. This raises concerns that the enthusiasm generated by such conventional evaluation metrics might mislead clinicians into incorrectly applying machine learning models in practice [59, 68].

In response, thus, we further annotated the FPR and FNR by varying the classification thresholds across the overall ROCs, ensuring clinicians better grasp the trade-off between different types of classification errors. For example, by setting a tolerance for FPR, they could assess whether the corresponding FNR is clinically acceptable. Such an analysis provides a clearer understanding of the model's practical relevance and its potential implications in real-world settings. Moreover, we also highlighted the results measured by Area under the precision-recall curve (AUPRC) that could overcome the optimism of AUROC [69], as it only focuses on the positive samples and will not be affected by the majority class of true negatives in disease detection. AUPRC summarizes the performance of a classifier across different thresholds by plotting precision against recall at varying thresholds. Our best time-independent RF detector achieved 0.35 ± 0.10 of AUPRC with a baseline (random guess) of 0.12 ± 0.02; and the shallow CNN detectors obtained approximately 0.38 ± 0.10 of AUPRC with a baseline of 0.13 ± 0.02. Although the values appear modest, the increased difference

from the baseline indicates a good performance. We could not directly compare the results with other studies as it was rarely reported in the literature.

Another critical aspect of the clinical utility concerns the process of model development and evaluation itself. Recent work has underscored the importance of properly framing the development and evaluation of models to assess their actual potential for clinical impact, as strong performance does not necessarily equate to clinically usable [70]. Some proposed ML algorithms can be biased when trained on variables that implicitly carry prior clinical knowledge. For example, including laboratory measurements in the sepsis prediction model may introduce bias, as the clinician's decision to order such tests or draw blood often reflects a pre-existing suspicion of disease [62–64].

Kamran et al. [36] further explored this issue by evaluating the discriminative ability of the widely implemented Epic sepsis model (Epic Systems Corporation, Verona, Wisconsin) in United States hospitals. They analyzed data from adult inpatients at the University of Michigan's academic medical center between October 2018 and December 2020 and compared the model's performance when making predictions before sepsis criteria were met and before clinical recognition (i.e., any treatment and diagnostic orders such as intravenous fluids, antibiotics, lactate measurement and blood culture were placed). The study revealed that the model had an AUROC of 0.62 (95% CI, 0.61 to 0.63) when calculating with predictions before sepsis criteria were met, but its performance dropped to an AUROC of 0.47 (95% CI, 0.46 to 0.48) when predictions were restricted to before any clinical indication was taken —essentially no better than random. This decline indicated that the model's apparent success was largely due to predictions made after clinicians had already identified and begun treating sepsis, raising questions about its ability to detect sepsis independently of clinical recognition.

A similar evaluation on a national-scale database in the United States was carried out by Beaulieu-Jones et al. [71]. They reported that ML models trained solely on clinician-initiated administrative data performed nearly as well as those trained on richer, more detailed EHR data. These findings suggest that many current ML models seem merely "looking over the shoulders of clinicians", offering little beyond what is already known through clinical expertise and judgment. In contrast, truly impactful ML models should "stand on the shoulders of clinicians", providing new insights and predictive power that clinicians alone cannot achieve.

These evaluations prompt researchers and healthcare professionals to rethink the genuine clinical utility of AI models, which remains a critical challenge. A model that largely relies on human intuition may lead to risks of overestimating its real-world applicability and can also result in misleading results when applied in different clinical settings where the timing and context of interventions vary. Moreover, a model that mimics clinical intuition but fails to offer novel, actionable insights risks contributing to alert fatigue rather than enhancing patient care [72]. As noted in [71], "*Machine learning can help clinicians make individualized patient predictions only if researchers demon-*

strate models that contribute novel insights, rather than simply predicting the next step a clinician will take." Therefore, clarifying the problem framework on which the model is built [70], ensuring it provides genuine added value, and avoiding over-reliance on data already influenced by clinical decisions are critical for ensuring the meaningful application of AI in healthcare.

5.6.4 Limitations and perspectives

This study has the following limitations and areas for improvement. One limitation lies in the data preparation and signal pre-processing. This study marks the first large-scale manipulation of the prospective, multi-center CARESS-Premi database, which contains about 30 years of accumulated monitoring signals, with an average of 20 days of signal acquisition per patient. Instead of focusing solely on sepsis-relevant episodes and patients, we processed all available raw signals throughout the entire follow-up periods of all enrolled patients. Given the volume of high-resolution data, e.g., ECG signals sampled at 500 Hz, we segmented the continuous signals into 6-hour windows for further processing. This process was driven by practical considerations of computational resources and limited time, but it resulted in a relatively low temporal resolution for sepsis detection, as models were trained on features extracted every 6 hours.

Although the use of 6-hour windows is consistent with previous studies [13, 17], it may be too coarse to capture the rapid changes that typically occur during the progression of sepsis. Sepsis can deteriorate quickly, and critical physiological changes may go undetected with this long window size. Recent studies have opted for finer-grained approaches, analyzing data in 3-hour [33], 1-hour [22, 29, 32, 58], 30-minute windows [34, 73], or even shorter time segments [65] to capture more dynamic fluctuations in vital signs or HRV parameters, which are crucial for early detection of sepsis. So in future work, adopting a finer temporal resolution, e.g., reducing the window size to 1 hour or less, may help enhance sepsis detection by capturing rapid, subtle changes in patient condition more accurately.

Another limitation concerns the framing of the problem, the importance of framing was thoroughly discussed in [70]. A common and useful manner adopted in almost all the studies on early sepsis detection is to play tricks on advancing positive labels to expected time windows before the defined onset, approaching the task as a binary classification task, so that the "classifications" for time windows before defined sepsis onsets would be equivalent to "predictions" for the clinical onsets. In contrast, in this study, as a starting point, we mainly focused on "sepsis detection" at the current moment, rather than anticipating future onset. This approach aims to determine whether an infant is currently at risk of developing sepsis, without projecting forward to future time points. While this framing allows us to model real-time sepsis detection, it potentially limits the model's predictive power in providing early warnings that could enable more timely interventions. Future work may benefit from combining real-time detection with predictive models to better balance immediate clinical needs with forward-looking risk predictions. This hybrid approach could improve

both the accuracy and utility of sepsis detection in NICU settings.

Within the framework of the CARESS-Premi project, our future research will extend beyond ECG signals to include chest impedance (respiratory signals) and plethysmography signals, which are already available in the dataset. These additional data streams offer a rich source of multi-modal physiological information, which may significantly enhance the accuracy of sepsis detection and early intervention models. By integrating multi-dimensional data, we will aim to develop models that combine continuous cardio-respiratory monitoring with advanced machine learning algorithms to improve early detection of sepsis and other critical conditions in neonates.

Moreover, a key focus moving forward will be on the dissemination, implementation, and validation of these predictive models across broader clinical settings [74]. Ensuring that models built on continuous cardiopulmonary data are tested in diverse, real-world environments is crucial for demonstrating their reliability and effectiveness. These efforts will include collaboration with healthcare professionals to streamline the interpretation of such data, along with refining the models based on real-time feedback from clinical applications. Ultimately, this work will aim to standardize the use of continuous monitoring data in the NICU and expand the applicability of our models to other conditions that benefit from early detection through real-time physiological monitoring.

5.7 Conclusion

In this chapter, we targeted a rather challenging issue of early detection of late-onset sepsis in preterm infants in the NICU setting. This topic has been largely studied in previous research, however, it remains a critical challenge. Here, we explored a non-invasive, computer-assisted approach to the early detection of neonatal LOS in preterm infants using HRV data derived from real-life monitoring in the NICU. Through the application of advanced signal processing techniques and machine learning models, we aimed to detect the presence of LOS and identify specific HRV patterns that precede clinical diagnosis, offering a tool to enhance the timeliness and accuracy of sepsis detection.

This study was conducted on a large multi-center cohort, after applying a series of inclusion and exclusion criteria, approximately 400 patients were included in the final analysis. A casual clinical timeline of infection was first formalized with time constants estimated from the literature, clinical expertise, or directly from our CARESS-Premi database. This formalization eases the representation of the causal effects that are involved during the LOS decision-making and allows us to propose an original approach in this field. By processing real-life, continuous, lasting-for-weeks, high-resolution monitoring ECG signals with an advanced signal processing chain, we extracted several HRV parameters that proved to be interesting for neonatal sepsis detection. On the other hand, we proposed a specific labeling strategy for longitudinal data with intertwined clinical

events. These procedures facilitated the dataset construction for machine learning model development. In terms of the ML algorithm, we not only used classic supervised learning for classification but also cleverly utilized the advantages of convolutional networks in processing time series to develop shallow CNN models. Results demonstrated that machine learning models trained on non-invasive HRV features and clinical information extracted from real-life monitoring data can effectively assist in the early detection of neonatal LOS. While the models achieved only modest accuracy, this study highlights the feasibility of leveraging HRV data for real-world applications in neonatal sepsis detection. Despite some limitations as discussed above, this work contributes to the ongoing development of predictive models in neonatal care, paving the way for more informed, data-driven, bedside clinical decisions in the management of neonatal sepsis.

BIBLIOGRAPHY

[1] A. L. Shane, P. J. Sánchez, and B. J. Stoll, "Neonatal sepsis," The lancet, vol. 390, no. 10104, pp. 1770–1780, 2017.

[2] R. P. Dellinger, M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke et al., "Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012," Critical care medicine, vol. 41, no. 2, pp. 580–637, 2013.

[3] J. Vincent, "The clinical challenge of sepsis identification and monitoring," PLoS medicine, vol. 13, no. 5, p. e1002022, 2016.

[4] V. S. Kuppala, J. Meinzen-Derr, A. L. Morrow, and K. R. Schibler, "Prolonged initial empirical antibiotic treatment is associated with adverse outcomes in premature infants," The Journal of pediatrics, vol. 159, no. 5, pp. 720–725, 2011.

[5] R. Singh, L. Sripada, and R. Singh, "Side effects of antibiotics during bacterial infection: mitochondria, the main target in host cell," Mitochondrion, vol. 16, pp. 50–54, 2014.

[6] A. Zea-Vera and T. J. Ochoa, "Challenges in the diagnosis and management of neonatal sepsis," Journal of tropical pediatrics, vol. 61, no. 1, pp. 1–13, 2015.

[7] World Health Organization, "Antimicrobial resistance," https://www.who.int/en/news-room/fact-sheets/detail/antimicrobial-resistance, November 2023, Accessed: 2024-09-30.

[8] I. H. Celik, M. Hanna, F. E. Canpolat, and M. Pammi, "Diagnosis of neonatal sepsis: the past, present and future," Pediatric research, vol. 91, no. 2, pp. 337–350, 2022.

[9] J. M. Sandell and I. K. Maconochie, "Paediatric early warning systems (PEWS) in the ED," pp. 754–755, 2016.

[10] B. Y. H. Wee, J. H. Lee, Y. H. Mok, and S.-L. Chong, "A narrative review of heart rate and variability in sepsis," Annals of translational medicine, vol. 8, no. 12, 2020.

[11] M. P. Griffin and J. R. Moorman, "Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis," Pediatrics, vol. 107, no. 1, pp. 97–104, 2001.

[12] D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman, "Sample entropy analysis of neonatal heart rate variability," American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, vol. 283, no. 3, pp. R789–R797, 2002.

[13] M. P. Griffin, T. M. O'Shea, E. A. Bissonette, F. E. Harrell, D. E. Lake, and J. R. Moorman, "Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness," Pediatric research, vol. 53, no. 6, pp. 920–926, 2003.

[14] B. P. Kovatchev, L. S. Farhy, H. Cao, M. P. Griffin, D. E. Lake, and J. R. Moorman, "Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome," Pediatric research, vol. 54, no. 6, pp. 892–898, 2003.

[15] M. P. Griffin, T. M. O'Shea, E. A. Bissonette, F. E. Harrell, D. E. Lake, and J. R. Moorman, "Abnormal heart rate characteristics are associated with neonatal mortality," Pediatric research, vol. 55, no. 5, pp. 782–788, 2004.

[16] H. Cao, D. E. Lake, M. P. Griffin, and J. R. Moorman, "Increased nonstationarity of neonatal heart rate before the clinical diagnosis of sepsis," Annals of biomedical engineering, vol. 32, pp. 233–244, 2004.

[17] M. P. Griffin, D. E. Lake, and J. R. Moorman, "Heart rate characteristics and laboratory tests in neonatal sepsis," Pediatrics, vol. 115, no. 4, pp. 937–941, 2005.

[18] M. P. Griffin, D. E. Lake, E. A. Bissonette, F. E. Harrell Jr, T. M. O'Shea, and J. R. Moorman, "Heart rate characteristics: novel physiomarkers to predict neonatal infection and death," Pediatrics, vol. 116, no. 5, pp. 1070–1074, 2005.

[19] J. R. Moorman, D. E. Lake, and M. P. Griffin, "Heart rate characteristics monitoring for neonatal sepsis," IEEE Transactions on Biomedical Engineering, vol. 53, no. 1, pp. 126–132, 2005.

[20] M. P. Griffin, D. E. Lake, T. M. O'Shea, and J. R. Moorman, "Heart rate characteristics and clinical signs in neonatal sepsis," Pediatric research, vol. 61, no. 2, pp. 222–227, 2007.

[21] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," American journal of physiology-heart and circulatory physiology, vol. 278, no. 6, pp. H2039–H2049, 2000.

[22] J. R. Moorman, W. A. Carlo, J. Kattwinkel, R. L. Schelonka, P. J. Porcelli, C. T. Navarrete, E. Bancalari, J. L. Aschner, M. W. Walker, J. A. Perez et al., "Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial," The Journal of pediatrics, vol. 159, no. 6, pp. 900–906, 2011.

[23] K. D. Fairchild, R. L. Schelonka, D. A. Kaufman, W. A. Carlo, J. Kattwinkel, P. J. Porcelli, C. T. Navarrete, E. Bancalari, J. L. Aschner, M. Walker et al., "Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial," Pediatric research, vol. 74, no. 5, pp. 570–575, 2013.

[24] L. Rio, A.-S. Ramelet, P. Ballabeni, C. Stadelmann, S. Asner, and E. Giannoni, "Monitoring of heart rate characteristics to detect neonatal sepsis," Pediatric Research, vol. 92, no. 4, pp. 1070–1074, 2022.

[25] S. Ahmad, A. Tejuja, K. D. Newman, R. Zarychanski, and A. J. Seely, "Clinical review: a review and analysis of heart rate variability and the diagnosis and prognosis of infection," Critical Care, vol. 13, pp. 1–7, 2009.

[26] E. Persad, K. Jost, A. Honoré, D. Forsberg, K. Coste, H. Olsson, S. Rautiainen, and E. Herlenius, "Neonatal sepsis prediction through clinical decision support algorithms: a systematic review," Acta Paediatrica, vol. 110, no. 12, pp. 3201–3226, 2021.

[27] S. Latremouille, J. Lam, W. Shalish, and G. Sant'Anna, "Neonatal heart rate variability: a contemporary scoping review of analysis methods and clinical applications," BMJ open, vol. 11, no. 12, p. e055209, 2021.

[28] F. J. Bohanon, A. A. Mrazek, M. T. Shabana, S. Mims, G. L. Radhakrishnan, G. C. Kramer, and R. S. Radhakrishnan, "Heart rate variability analysis is more sensitive at identifying neonatal sepsis than conventional vital signs," The American Journal of Surgery, vol. 210, no. 4, pp. 661–667, 2015.

[29] L. Cabrera-Quiros, D. Kommers, M. K. Wolvers, L. Oosterwijk, N. Arents, J. van der Sluijs-Bens, E. J. E. Cottaar, P. Andriessen, and C. van Pul, "Prediction of late-onset sepsis in preterm infants using monitoring signals and machine learning," Critical Care Explorations, vol. 3, no. 1, p. e0302, 2021.

[30] Z. Peng, G. Varisco, X. Long, R.-H. Liang, D. Kommers, W. Cottaar, P. Andriessen, and C. van Pul, "A continuous late-onset sepsis prediction algorithm for preterm infants using multi-channel physiological signals from a patient monitor," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 1, pp. 550–561, 2023.

[31] A. Beuchée, G. Carrault, J. Y. Bansard, E. Boutaric, P. Bétrémieux, and P. Pladys, "Uncorrelated randomness of the heart rate is associated with sepsis in sick premature infants," Neonatology, vol. 96, no. 2, pp. 109–114, 2009.

[32] M. Yang, Z. Peng, C. van Pul, P. Andriessen, K. Dong, D. Silvertand, J. Li, C. Liu, and X. Long, "Continuous prediction and clinical alarm management of late-onset sepsis in preterm infants using vital signs from a patient monitor," Computer Methods and Programs in Biomedicine, p. 108335, 2024.

[33] R. Joshi, D. Kommers, L. Oosterwijk, L. Feijs, C. van Pul, and P. Andriessen, "Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics, and ECG-derived estimates of infant motion," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 3, pp. 681–692, 2020.

[34] C. León, G. Carrault, P. Pladys, and A. Beuchée, "Early detection of late onset sepsis in premature infants using visibility graph analysis of heart rate variability," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 4, pp. 1006–1017, 2021.

[35] F. Brunkhorst, U. Heinz, and Z. Forycki, "Kinetics of procalcitonin in iatrogenic sepsis," Intensive care medicine, vol. 24, no. 8, pp. 888–889, 1998.

[36] F. Kamran, D. Tjandra, A. Heiler, J. Virzi, K. Singh, J. E. King, T. S. Valley, and J. Wiens, "Evaluation of sepsis prediction models before onset of treatment," NEJM AI, vol. 1, no. 3, p. AIoa2300032, 2024.

[37] P. Gastmeier, C. Geffers, F. Schwab, J. Fitzner, M. Obladen, and H. Rüden, "Development of a surveillance system for nosocomial infections: the component for neonatal intensive care units in germany," Journal of Hospital Infection, vol. 57, no. 2, pp. 126–131, 2004.

[38] M. Doyen, D. Ge, A. Beuchée, G. Carrault, and A. I. Hernández, "Robust, real-time generic detector based on a multi-feature probabilistic method," PLoS ONE, vol. 14, no. 10, p. e0223785, 2019.

[39] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger et al., "Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology," Circulation, vol. 93, no. 5, pp. 1043–1065, 1996.

[40] A. Bauer, J. W. Kantelhardt, A. Bunde, P. Barthel, R. Schneider, M. Malik, and G. Schmidt, "Phase-rectified signal averaging detects quasi-periodicities in non-stationary data," Physica A: Statistical Mechanics and its Applications, vol. 364, pp. 423–434, 2006.

[41] D. R. Kommers, R. Joshi, C. van Pul, L. Atallah, L. Feijs, G. Oei, S. B. Oetomo, and P. Andriessen, "Features of heart rate variability capture regulatory changes during kangaroo care in preterm infants," The journal of pediatrics, vol. 182, pp. 92–98, 2017.

[42] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," Frontiers in public health, vol. 5, p. 258, 2017.

[43] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," Chaos: an interdisciplinary journal of nonlinear science, vol. 5, no. 1, pp. 82–87, 1995.

[44] T. Nakamura, H. Horio, S. Miyashita, Y. Chiba, and S. Sato, "Identification of development and autonomic nerve activity from heart rate variability in preterm infants," Biosystems, vol. 79, no. 1-3, pp. 117–124, 2005.

[45] J. L. Wynn, S. O. Guthrie, H. R. Wong, P. Lahni, R. Ungaro, M. C. Lopez, H. V. Baker, and L. L. Moldawer, "Postnatal age is a critical determinant of the neonatal host response to sepsis," Molecular Medicine, vol. 21, pp. 496–504, 2015.

[46] C. Vásquez, A. Hernández, F. Mora, G. Carrault, and G. Passariello, "Atrial activity enhancement by wiener filtering using an artificial neural network," IEEE Transactions on Biomedical Engineering, vol. 48, no. 8, pp. 940–944, 2001.

[47] F. Porée, A. Kachenoura, G. Carrault, R. Dal Molin, P. Mabo, and A. I. Hernández, "Surface electrocardiogram reconstruction from intracardiac electrograms using a dynamic time delay artificial neural network," IEEE Transactions on Biomedical Engineering, vol. 60, no. 1, pp. 106–114, 2012.

[48] S. Ioffe, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[50] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in NIPS-W, 2017.

[51] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," Technometrics, vol. 33, no. 2, pp. 161–174, 1991.

[52] J. L. Wynn, H. R. Wong, T. P. Shanley, M. J. Bizzarro, L. Saiman, and R. A. Polin, "Time for a neonatal-specific consensus definition for sepsis," Pediatric Critical Care Medicine, vol. 15, no. 6, pp. 523–528, 2014.

[53] J. L. Wynn and R. A. Polin, "Progress in the management of neonatal sepsis: the importance of a consensus definition," Pediatric Research, vol. 83, no. 1, pp. 13–15, 2018.

[54] ——, "A neonatal sequential organ failure assessment score predicts mortality to late-onset sepsis in preterm very low birth weight infants," Pediatric research, vol. 88, no. 1, pp. 85–90, 2020.

[55] M. McGovern, E. Giannoni, H. Kuester, M. A. Turner, A. van Den Hoogen, J. M. Bliss, J. M. Koenig, F. M. Keij, J. Mazela, R. Finnegan et al., "Challenges in developing a consensus definition of neonatal sepsis," Pediatric research, vol. 88, no. 1, pp. 14–26, 2020.

[56] World Health Organization, "Vermont Oxford Network: Manual of Operations: Part 2-data definitions & infant data forms. Release 28.0," https://https://vtoxford.zendesk.com/hc/en-us/articles/21280617139859-2024-Manual-of-Operations-Part-2-Release-28-0-PDF, February 2024, Accessed: 2024-10-14.

[57] European Medicines Agency (EMA) London, "Report on the expert meeting on neonatal and paediatric sepsis," https://www.ema.europa.eu/en/documents/report/report-expert-meeting-neonatal-and-paediatric-sepsis_en.pdf, June 2010, Accessed: 2024-10-14.

[58] M. A. van den Berg, O. O'Jay, M. M. Benders, R. R. Bartels, D. D. Vijlbrief et al., "Development and clinical impact assessment of a machine-learning model for early prediction of late-onset sepsis," Computers in Biology and Medicine, vol. 163, p. 107156, 2023.

[59] M. Meeus, C. Beirnaert, L. Mahieu, K. Laukens, P. Meysman, A. Mulder, and D. Van Laere, "Clinical decision support for improved neonatal care: The development of a machine learning model for the prediction of late-onset sepsis and necrotizing enterocolitis," The Journal of Pediatrics, vol. 266, p. 113869, 2024.

[60] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, and A. Sharma, "Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019," Critical care medicine, vol. 48, no. 2, pp. 210–217, 2020.

[61] M. Chen and A. Hernández, "Towards an explainable model for sepsis detection based on sensitivity analysis," IRBM, vol. 43, no. 1, pp. 75–86, 2022.

[62] S. Mani, A. Ozdas, C. Aliferis, H. A. Varol, Q. Chen, R. Carnevale, Y. Chen, J. Romano-Keeler, H. Nian, and J.-H. Weitkamp, "Medical decision support using machine learning for early detection of late-onset neonatal sepsis," Journal of the American Medical Informatics Association, vol. 21, no. 2, pp. 326–336, 2014.

[63] A. J. Masino, M. C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C. P. Bonafide, F. Balamuth, M. Schmatz, and R. W. Grundmeier, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," PloS one, vol. 14, no. 2, p. e0212665, 2019.

[64] A. C. Helguera-Repetto, M. D. Soto-Ramírez, O. Villavicencio-Carrisoza, S. Yong-Mendoza, A. Yong-Mendoza, M. León-Juárez, J. A. González-y Merchand, V. Zaga-Clavellina, and C. Irles, "Neonatal sepsis diagnosis decision-making based on artificial neural networks," Frontiers in pediatrics, vol. 8, p. 525, 2020.

[65] A. Honoré, D. Forsberg, K. Adolphson, S. Chatterjee, K. Jost, and E. Herlenius, "Vital sign-based detection of sepsis in neonates using machine learning," Acta Paediatrica, vol. 112, no. 4, pp. 686–696, 2023.

[66] S. A. Coggins, J.-H. Weitkamp, L. Grunwald, A. R. Stark, J. Reese, W. Walsh, and J. L. Wynn, "Heart rate characteristic index monitoring for bloodstream infection in an NICU: a 3-year experience," Archives of Disease in Childhood-Fetal and Neonatal Edition, vol. 101, no. 4, pp. F329–F332, 2016.

[67] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4. Springer, 2009, pp. 461–471.

[68] J. Wilkinson, K. F. Arnold, E. J. Murray, M. van Smeden, K. Carr, R. Sippy, M. de Kamps, A. Beam, S. Konigorski, C. Lippert et al., "Time to reality check the promises of machine learning-powered precision medicine," The Lancet Digital Health, vol. 2, no. 12, pp. e677–e680, 2020.

[69] B. Ozenne, F. Subtil, and D. Maucort-Boulch, "The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," Journal of clinical epidemiology, vol. 68, no. 8, pp. 855–859, 2015.

[70] S. M. Lauritsen, B. Thiesson, M. J. Jørgensen, A. H. Riis, U. S. Espelund, J. B. Weile, and J. Lange, "The framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards," NPJ digital medicine, vol. 4, no. 1, p. 158, 2021.

[71] B. K. Beaulieu-Jones, W. Yuan, G. A. Brat, A. L. Beam, G. Weber, M. Ruffin, and I. S. Kohane, "Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?" NPJ digital medicine, vol. 4, no. 1, p. 62, 2021.

[72] J. J. Cash, "Alert fatigue," American Journal of Health-System Pharmacy, vol. 66, no. 23, pp. 2098–2101, 2009.

[73] A. G. Garstman, C. Rodriguez Rivero, and W. Onland, "Early detection of late onset sepsis in extremely preterm infants using machine learning: Towards an early warning system," Applied Sciences, vol. 13, no. 16, p. 9049, 2023.

[74] S. B. Walker, C. M. Badke, M. S. Carroll, K. S. Honegger, A. Fawcett, D. E. Weese-Mayer, and L. N. Sanchez-Pinto, "Novel approaches to capturing and using continuous cardiorespiratory physiological data in hospitalized children," Pediatric research, vol. 93, no. 2, pp. 396–404, 2023.

CHAPTER

6 Deployment of an On-the-Edge Clinical Decision Support System in Neonatal Intensive Care Units

In recent decades, Clinical decision support systems (CDSS) have emerged as valuable technologies for managing complex medical conditions. A range of CDSS applications are focused on neonatal care, including the management of hyperbilirubinemia, medication, nutrition optimization, and risk estimation for morbidity, mortality, and sepsis [1], such as Artemis [2], Baby Steps [3], Etiometry [4, 5] and iNICU [6, 7]. With the progress in research and the evolution of technical resources within NICU, establishing an AI-based clinical decision support systems that facilitate early intervention for preterm neonates is challenging yet promising.

However, although the development and deployment of CDSS integrated into hospitals and intensive care units gained increasing attention, there are rare cases of on-the-edge CDSS that have been validated technically and clinically in real-life medical scenarios. Given the surge in medical data, resource limitations, and the need for real-time, low-latency, and bandwidth-efficient processing, working on the edge is becoming an original and appealing direction.

In this chapter, we propose to design, implement, deploy, and technically evaluate a CDSS that integrates a signal processing chain and machine learning inference models on the edge in the scope of neonatology, taking the NICU in the *University Hospital Center of Rennes* (*CHU Rennes*) as a pilot, allowing for quasi-real-time processing and fusion of high-resolution monitoring time series and data in the Hospital Information System (HIS). After the architecture concepts, preliminary results of the proposed system concerning the technical performance are presented, followed by a simple use case of the baseline model developed in Chapter 4 to validate the clinical feasibility of the proposed system. Afterward, some implications for the early detection of neonatal sepsis, related to Chapter 5, based on the proposed system are presented in the form of a case report.

6.1 Introduction

Developing Clinical decision support systems (CDSS) that facilitate personalized medicine requires the use and linking of massive data from different sources. The temporal evolution of physiological data, which is the basis of clinical reasoning in intensive care units, is not yet commonly

included in data warehouses, although this appears essential in the medium term. The proposed system was recently introduced to the market with the aim of easily providing information on a daily basis, but also as a data source service for research, particularly in the field of learning methods on massive data (artificial intelligence). The data warehouse system (Patient Information Center iX, Data Warehouse Connect, Philips Medical Systems, Andover, MA) has been used in a few clinical studies [8–11]. The system enables the creation and utilization of comprehensive databases that integrate all monitoring data, in particular physiological signals, with sampling adapted for signal processing. It also allows the incorporation of monitoring parameters and data from connected peripheral devices (e.g., ventilation, perfusion systems, etc.). By leveraging the requested data warehouse system, it becomes possible to associate immediate clinical use with the development of personalized medicine solutions, such as the management of neonatal hyperbilirubinemia and neonatal sepsis, as discussed and developed in previous chapters.

However, any proposed CDSS must undergo technical validation in real-world conditions over a sufficient period of time before progressing to clinical validation through a multi-center randomized controlled study.

The on-the-edge CDSS, we propose, is derived from the concept of "edge computing", which deploys computing resources on the edge side near data sources and operates on "instant data" that are generated in real-time. Edge-based solutions provide the framework for reduced latency for time-dependent solutions, such as vital sign monitoring, and they offer added security during data transmission compared to traditional computing systems [12]. While there has been an impressive exploration of edge computing in smart healthcare systems, much of the focus has been on wearable devices, smartphone-based sensors and ambient applications [12]. There is, however, a notable gap in the development and deployment of edge-based CDSS integrated into hospital and ICU environments, where the collection and storage of critical medical data dominate the healthcare system in both volume and importance. Given the enormous amount of medical data generated every second, resource constraints, and the need for real-time data processing with low latency and reduced network bandwidth, working on the edge offers an attractive and innovative solution.

To achieve this, a robust technological architecture is required, based on the Data Warehouse Connect (Philips), capable of storing, analyzing and presenting high-resolution physiological data from monitoring monitors. Specifically, such a system should support: *i*) the recovery of sufficiently sampled signals from patient monitors, *ii*) the secure storage of data on a hospital clinical server, *iii*) integration with proposed clinical decision support systems, *iv*) secure interfacing with research infrastructures (*LTSI - INSERM U1099*), and *v*) the accessibility for both the clinical and research teams to analyze and evaluate the collected and stored data after anonymization.

6.2 System Architecture

Before detailing our proposed system, it is essential to first introduce the monitoring environment in neonatal intensive care settings.

6.2.1 Monitoring environment in NICU

The proposed system is designed to integrate seamlessly into the existing infrastructure at the NICU within *CHU Rennes*, where the installed patient monitoring and data management solution are from Philips (Eindhoven, Netherlands). As shown in Figure 6.1, the medical data flow in this setup can be summarized as three main stages: data acquisition, aggregation and storage & distribution.

At the patient's bedside, monitors (IntelliVue MP40) and integration solutions (IntelliBridge System) capture and transmit vital monitoring data. This data is then sent to the Patient Information Center iX (PIC iX) system, a core part of the Philips enterprise monitoring ecosystem, which manages real-time patient monitoring and provides a comprehensive view of patient conditions by integrating data from both Philips and non-Philips devices.

Data Warehouse Connect (DWC) is a licensed feature of the PIC iX system that allows storage of high-resolution data from Philips patient monitoring devices, telemetry devices, and third-party devices connected to the IntelliBridge family of products.

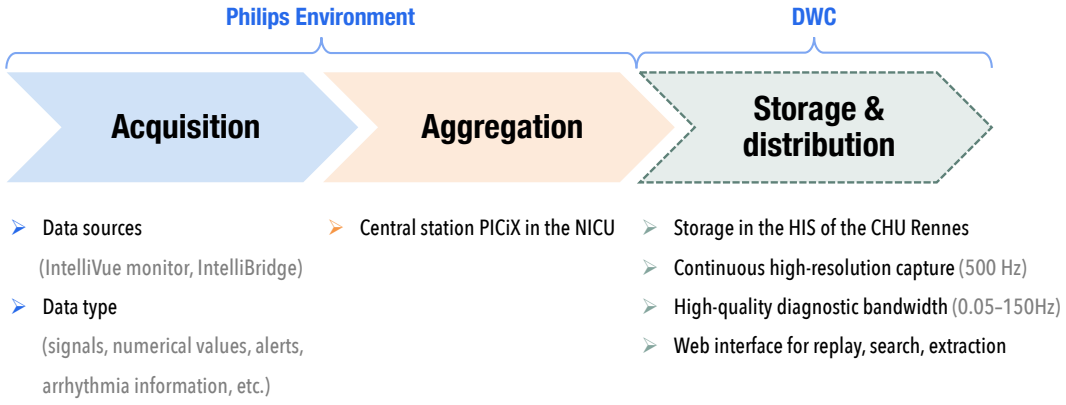

Figure 6.1: Integration into the infrastructure in the NICU.

6.2.2 Proposed on-the-edge CDSS architecture

The proposed on-the-edge CDSS architecture consists of three main components, as illustrated in Figure 6.2:

1. NICU bedside monitors connected to Philips' Data Warehouse Connect (DWC);

2. Electronic health records (EHR) system acquiring data from other devices and the Hospital Information System (HIS);
3. A Virtual machine (VM) deployed into a restricted network for holding our CDSS.

The data flow within the proposed system follows two parallel pathways, indicated by the blue and yellow arrows in Figure 6.2. In the first pathway, as a standard care routine for neonates from birth to nearly discharge, high-resolution time series continuously monitored by bedside monitors, along with structured data (e.g., instant heart rates, annotations, alerts, etc.) computed by the monitors' built-in algorithms, are stored on the DWC server. The DWC writes this monitoring data to a Philips-implemented custom SQL agent every 5 seconds. This SQL agent is configured to automatically export the data from the previous 20 to the previous 5 minutes every 15 minutes, with a 5-minute delay due to transmission and storage time consumed from the bedside monitors to the DWC and the SQL database.

Simultaneously, the second pathway involves various medical devices that continuously collect clinical metadata, such as patient demographics, respiratory support data, nutritional support, laboratory test results and clinical events. All information is managed by the HIS (MetaVision, iMDsoft, Tel Aviv, Israel) and stored in the EHR, with updates recorded at a lower resolution (1 data point per minute). This metadata is selectively exported every 10 minutes.

Both data streams undergo pseudonymization using a common ID before the data segments are copied to a shared mount point (**IN**).

The processing unit, as shown on the right side of Figure 6.2, is activated on a first-in-first-out (FIFO) basis once a new batch of data arrives at the **IN** (i.e., every 15 minutes). It allows quasi-real-time processing of high-resolution time series using ML algorithms encapsulated within a VM. Once processed, the data segments are erased to free up storage, and the output is written to another shared point (**OUT**) at the end of the workflow. The results stored in **OUT** are designed to be returned to caregivers in the electronic health record through the HL7 communication protocol or via a dedicated decision support system [13]. These results, presented as numerical data and waveforms, provide essential insights to help caregivers adjust clinical interventions promptly.

6.2.3 Data specification for IN and OUT

IN format

At the **IN** mount point, the metadata exported by the EHR varies in format and content depending on the specific study. In our use case, the metadata of all available patients is compiled in a table with columns for patient ID (denoted as IEP, the hospital episode identifier provided by *CHU Rennes*), Postmenstrual age (PMA), Postnatal age (PNA) and Gestational age (GA). The IEP serves as a unique identifier linking to the signals data exported from the DWC. In cases where the IEP is not immediately available from the hospital, a temporary Study ID provided by the DWC

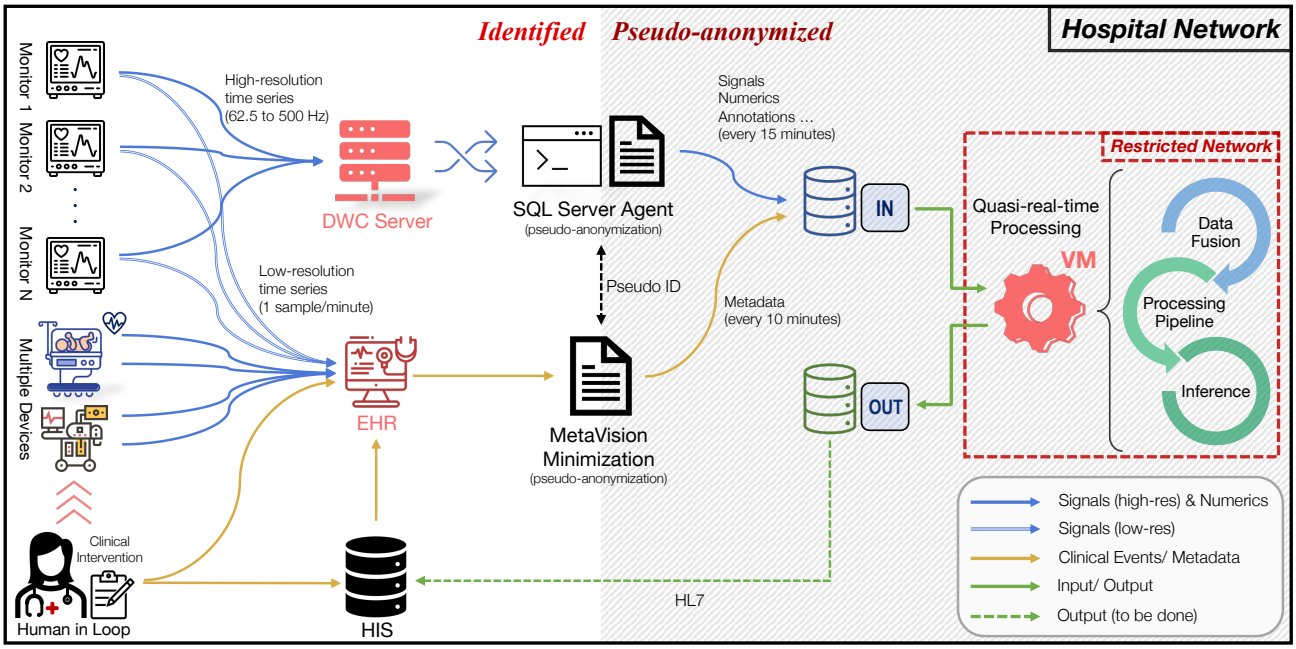

Figure 6.2: Architecture of the proposed on-the-edge system for quasi-real-time clinical decision support in NICU.

(the Philips side) is used. The IEP is set as *None* for the patient until the hospital provides it via metadata in subsequent data exports.

The files exported by the DWC are organized in folders named with patient IDs. Within each patient-specific folder, files are named using a combination of the patient ID, timestamp and file attributes, i.e., `PatientID_YYYYMMDDhhmmssms_<attribute>.<filetype>`.

The structured data includes numeric values (stored as "*numerics.csv*"), cardiac annotations ("*annotations.csv*") and patient identification ("*patient.id*"). Raw signals in MIT data format ("*.dat*") are also exported, but only once every 1 hour on request for the current version. A detailed overview of the naming convention is listed in Table 6.1, and an example of the file organization within the **IN** mount point is shown in Figure 6.3.

OUT format

For each patient, all results are grouped in an independent folder named after the patient ID. The folder contains "X.csv" (representing the input features for machine learning models) and "Y.csv" (representing the output of the model) both named with timestamps corresponding to the monitoring time. In addition, log files that document the execution status of the entire system are also saved in the **OUT** folder. An example of the file structure in the **OUT** mount point is provided in Figure 6.4.

Table 6.1: Naming convention of the files in the IN mount point.

Section	Description
PatientID	IEP (hospital episode identifier, 9 digits) StudyID (a sequential anonymous ID generated by DWC system)
YYYYMMDDhhmmssms	This is the timestamp of the start of the record: YYYY = year MM = month DD = day of month hh = hour of the day mm = minutes ss = seconds ms = milliseconds to 7 digits e.g. If start timestamp in the database was 2022-03-18 07:26:47.9950100, the presentation in the file name will be 202203180726479950100.
Attribute	Optional string representing the following possible content: numerics = numeric values annot = string values e.g. ventilator mode, ECG beat labels <wave label>= short label of the waveform, e.g., II, Resp, Pleth, EEG, etc.
Filetype	File extensions, identifying the following possible content: .hea = header text file of a segment of waveforms .clock = reference clock text file for a given .hea file .dat = waveform binary files .CSV = numeric and text values recorded at patient bedside .id = patient identification, first name and last name

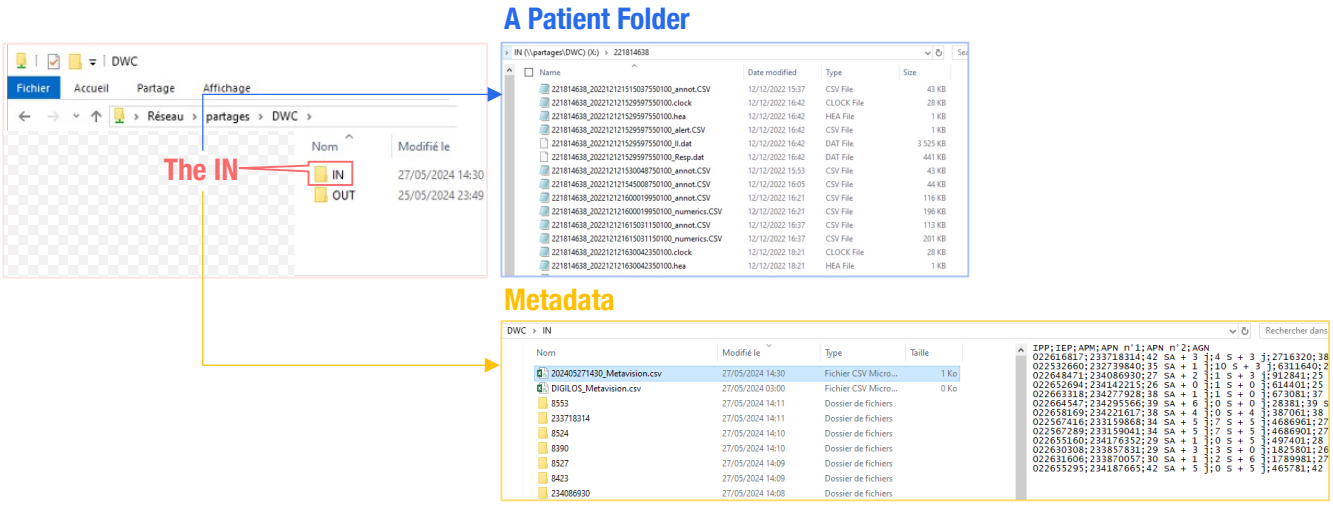

Figure 6.3: File organization in the IN mount point of the system.

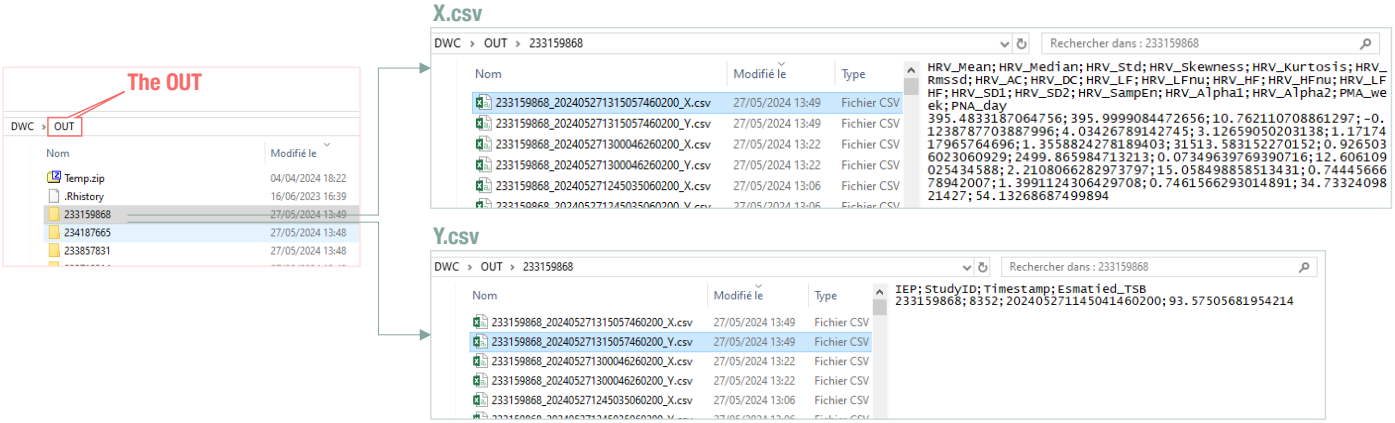

Figure 6.4: File organization in the OUT mount point of the system.

TEMP format

We also configure an additional **TEMP** folder to store intermediate results as backups. Each processed patient has a unique ".csv" file that includes associated metadata and extracted HRV parameters arranged in chronological order. These intermediate results might be useful for further offline analysis and model development.

6.2.4 Core processing unit

The system's core processing unit operates within a VM environment. A simplified diagram of the processing engine is presented in Figure 6.5, showcasing three key components: *i)* data fusion, *ii)* data processing, and *iii)* model inference. These components interact with two mount points for input and output: **IN** and **OUT**. A more detailed workflow of the actual implementation is illustrated in Figure 6.6.

The main process runs in a non-stop loop (*while True*) to perform a thorough scan of **IN** folder to summarize the incoming metadata and patient folders. Using a *for loop*, the core processing unit processes the data in the **IN** patient by patient, i.e., folder by folder. All processing activities are recorded in a series of log files during this execution.

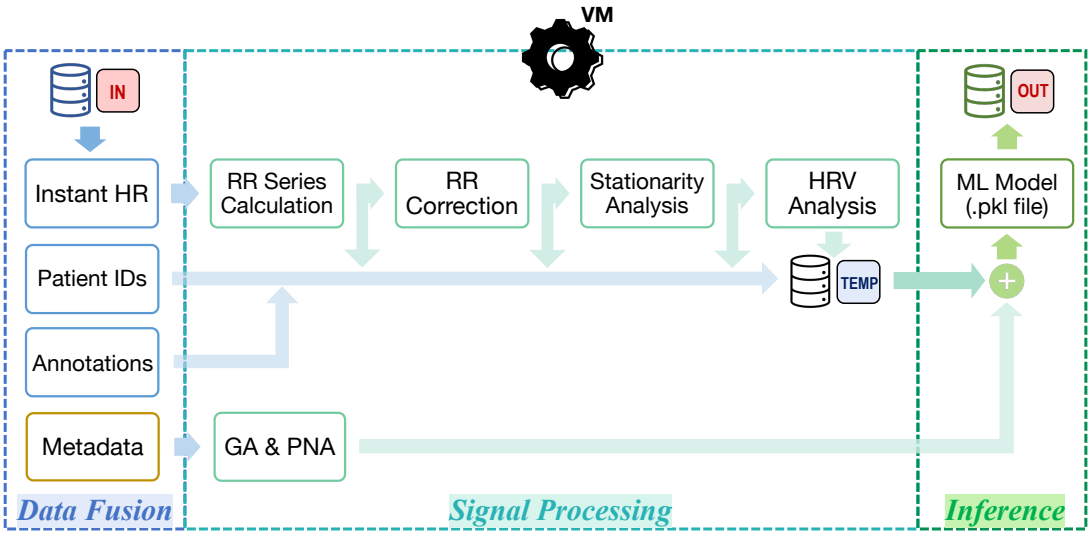

Figure 6.5: Simplified diagram of the core data processing unit.

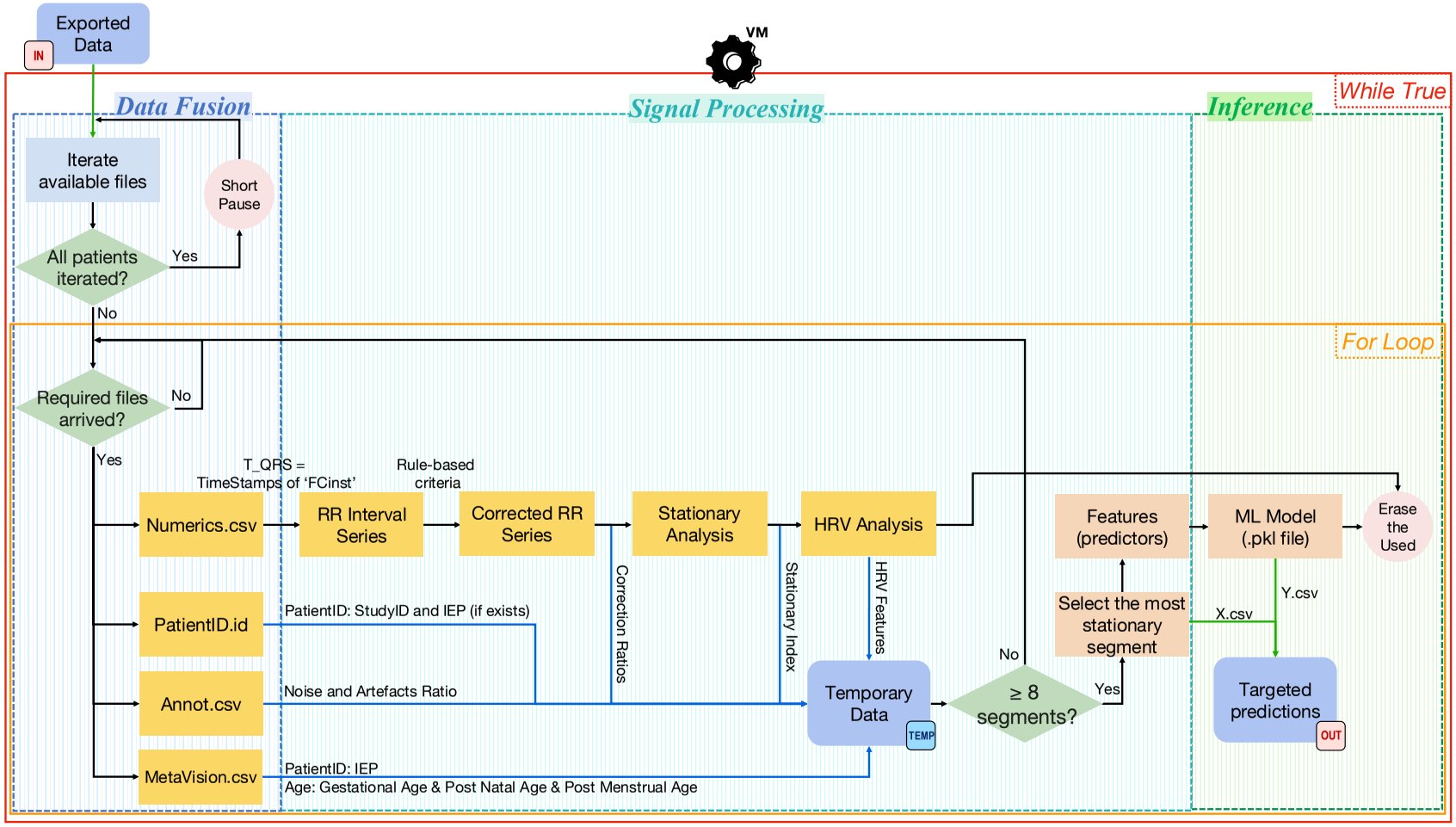

Figure 6.6: Detailed diagram of the core data processing unit.

The first step involves fusing data from two sources—DWC and EHR—by matching patient IDs and synchronizing timestamps. Before data can proceed to the next module, a readiness check ensures that all four necessary files, as described in the **IN format** section, have arrived in the patient folder, and none are empty. Each file should be named by a matching patient ID and a corresponding timestamp. The metadata file is parsed first to search for the relevant patient ID, and if found, the age information (PNA and GA) is selected and calibrated with the timestamp recorded in the filename for synchronization.

The next step is the signal processing pipeline that transforms multi-source structured data into clinically relevant features. In this first version of the deployment, instead of detecting cardiac beats (QRS complexes) from raw ECG signals, we directly exploit the instantaneous heart rates generated by Philips monitors to construct the RR interval series. A multi-step approach of logic rules based on pathology and rhythm correction is then applied to automatically detect and correct possible artifacts and errors in the RR series (as detailed in Section 2.2.3). Then alterations of the mean and variance of the corrected series over time are estimated, and signal stationarity is analyzed (detailed in Section 2.2.4). Afterward, a set of Heart rate variability (HRV) features, characterizing cardiovascular functions modulated by the autonomic nervous system, is extracted from the segments (as detailed in Section 2.2.5). These include:

- Twelve time-domain features:
  Mean, Median, Std, Rmssd, Skewness, Kurtosis, DC (deceleration capacity) and AC (acceleration capacity), the interdecile range between 10^{th} and 90^{th} percentile (IDR), percentage deceleration of RR intervals (pDec), the standard deviation of RR corresponding to pDec (stdDec) and sample asymmetry of RR histogram (SampAsym).
- Five frequency-domain features:
  Power in the low-frequency spectral band (LF, 0.02-0.2 Hz), normalized LF (LFnu), power in the high-frequency band (HF, 0.2-2 Hz), normalized HF (HFnu); and LF/HF ratio.
- Five non-linear features:
  SD1 and SD2 from the Poincaré plot, sample entropy (SampEn), and $\alpha 1$, $\alpha 2$ from the detrended fluctuation analysis.

During the data fusion and processing stages, a temporary file is created as a buffer to store all calculated intermediate results. This file is written into the **TEMP** folder.

The final stage concerns machine learning model inference. A previously trained ML estimator is activated in inference mode to generate predictions. Depending on specific tasks, the model takes extracted features as input ("X.csv") and inference the predictions to output ("Y.csv"), which are directly made available to clinicians. Importantly, the inference component/instance is designed as a plug-and-play plugin, compatible with any ML model in the PICKLE format (*.pkl*), allowing for easy generalization of the system for other inference applications.

During parsing and processing, any files with incorrect formats or erroneous data are auto-

matically erased from the **IN**, ensuring the system's robustness. It is also worth mentioning that we configured additional data quality validation checks throughout the processing chain to guarantee high Signal-to-noise ratio (SNR) and data reliability, especially given the complexity and heterogeneity of the data from real-life, long-term monitoring in the NICU.

6.3 System Performance: Preliminary technical validation

The technical objective of this study is to validate the feasibility and stability of the proposed architecture. To this end, we conducted a pilot study under real conditions over a sufficient period of time. The infrastructure was installed at *CHU Rennes* in collaboration with Philips and our laboratory (*LTSI - INSERM U1099*).

The system was successfully deployed and implemented at *CHU Rennes* since January 2023. As of June 2023 (preliminary testing of system working performance), the system has been continuously running for several months, constantly receiving data from a total of 138 authorized neonates born within the first six months of 2023 during their NICU stay. The median [IQR] GA of the neonates is 35.5 [30.3; 39.4] weeks, and their median [IQR] birth weight is 2,240 [1,341; 3,211] grams.

As an initial use case, we embedded a baseline Random Forest regressor (RF_{base}), previously trained for bilirubin level estimation (refer to Chapter 4), as the inference model into the system. The system processes live data, generating bilirubin estimations (continuous variables) for each infant with a 15-minute time resolution, while minimizing computation time and memory usage.

Table 6.2: Technical performance of the proposed on-the-edge CDSS.

	Memory Usage*	Execution Time†	System Latency†
Max.	206.4 MB	5.56 s	29.4 min
Min.	168.5 MB	0.04 s	20.1 min
Mean	187.0 MB	2.36 s	26.5 min
Std.	8.84 MB	0.45 s	1.95 min
Median	187.5 MB	2.40 s	26.8 min

*Statistics are calculated from June 9, 2023 to June 19, 2023.

†Statistics are calculated for the 24 hours from the noon of June 18, 2023 to the noon of June 19, 2023.

Table Table 6.2 summarizes the system's technical performance that is measured in given periods, focusing on memory usage, execution time, and system latency. As shown, the proposed VM-based system exhibits a low memory footprint, consuming merely 187.0±8.84 MB during execution. Execution time refers to the time taken for one patient's qualified data segment to completely

undergo quasi-real-time processing, starting from the data's arrival at the **IN** folder, through the processing pipeline, and ending with the output of the results, followed by memory release. During a 24-hour period, a total of 3,293 data segments were qualified and processed, with an average execution time of 2.36 ± 0.45 seconds.

System latency encompasses various parts, including acquisition delay (time taken for bedside monitor data to be transmitted to the DWC), database query delay (an inherent 20-minute delay due to the SQL agent configuration), transmission delay (from the SQL database to the **IN** mount point), processing delay (execution time), and waiting time (time spent waiting for resources within the system to be processed or responded to). The minimum overall latency observed in the considered period was 20.1 minutes, indicating the possibility of the whole processing being finalized immediately after the data arrival when subtracting the inherent 20 minutes.

Figure 6.7 visualizes the execution time and latency over a 24-hour time span. The x-axis represents the timeline, divided into 15-minute intervals, and the y-axis shows the response time in seconds for each data batch. In general, the majority of data segments processed and output to the **OUT** (represented by red scatter points) exhibit processing times between 2 and 3 seconds. Observing the timeline, the data arrives every 15 minutes as configured, and each data transmission and processing cycle is normally completed within 5 minutes. This indicates that the system has sufficient capacity to handle higher concurrency and throughput, which may accommodate more resource-intensive processing tasks in future deployments.

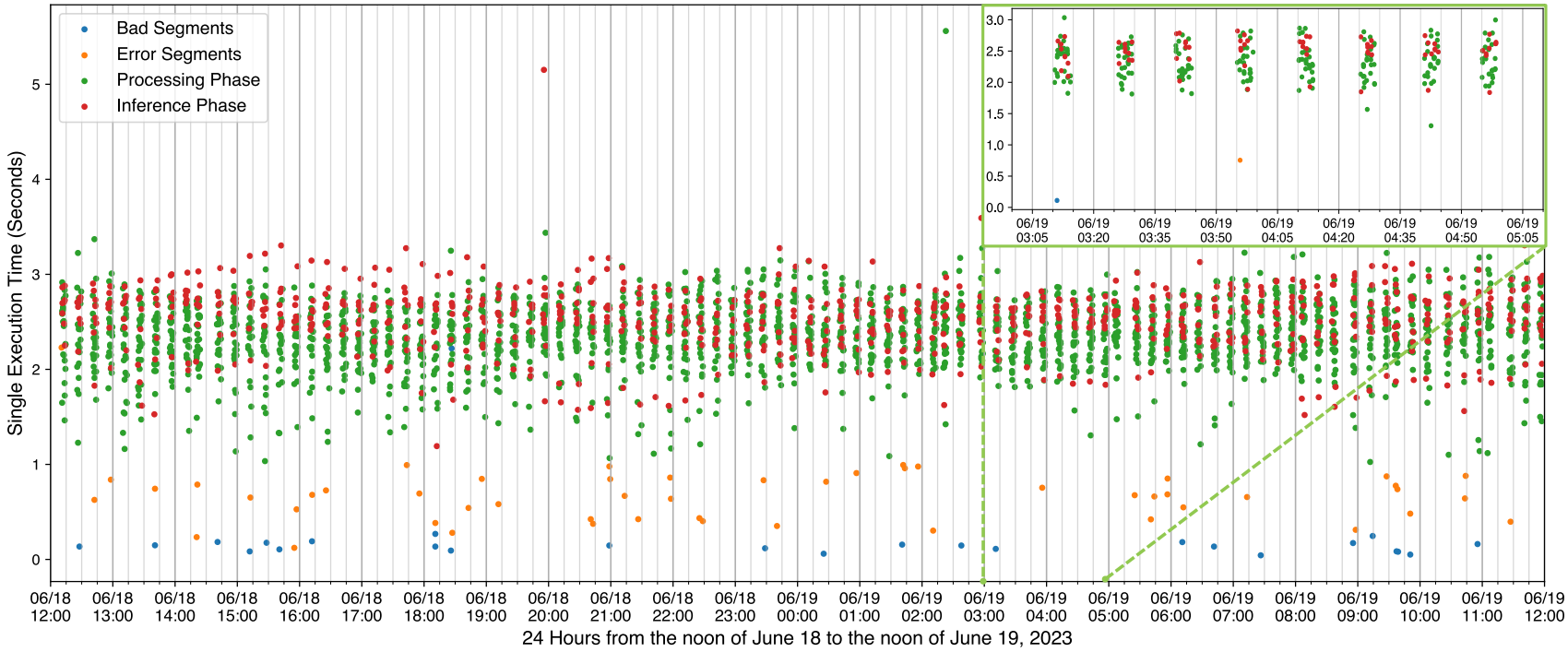

Figure 6.7: Visualization of execution time and system latency.

6.4 Case Report: Implications on early detection of neonatal sepsis from the on-the-edge system

As a proof-of-concept, the proposed on-the-edge clinical decision support system aims to integrate real-time clinical monitoring signals from DWC and electronic health information from the HIS and generate useful results through a VM-based processing engine that could benefit the optimization of neonatal outcomes. Regarding the detection of LOS, we have not deployed yet any inference model, but the VM keeps generating HRV features that are provided to the clinician. Here we present a brief case report showing the interest of proving such complementary information. Although we have not yet closed the loop on the entire system (returning the results to the HIS for all caregivers in the NICU to share), as a pilot study, the clinician responsible for the project has direct access to the **IN** and **OUT** mount points and the **TEMP** folder. Interesting implications are observed from the preliminary results.

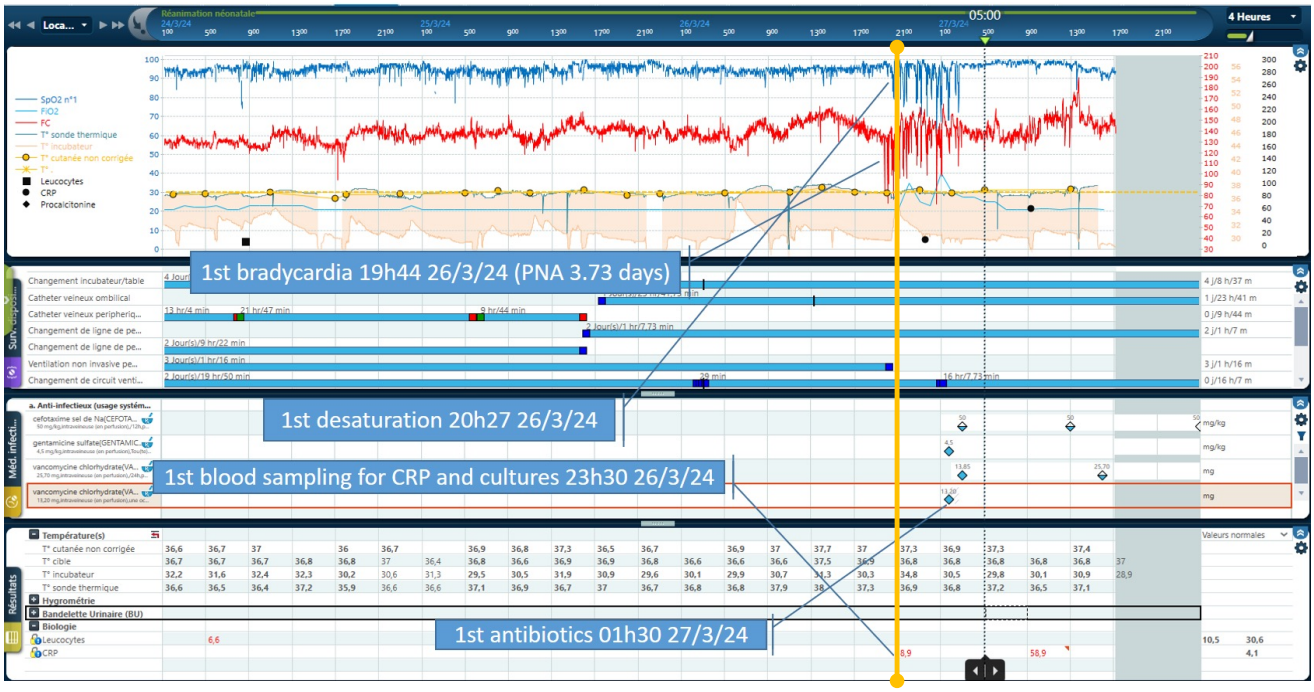

Figure 6.8: A clinical example of a delay between heart rate changes, blood sampling and antibiotics starting in an infant with clinically proven infection shown by MetaVision (the HIS installed in *CHU Rennes*).

Figure 6.8 displays a clinical example of the delay between heart rate changes, blood sampling and antibiotics starting in a preterm infant born on 23 March 2024 at 2:17 with a gestational age of $29^{5/7}$ weeks and a birth weight of 1,245 grams. The first upper panel shows the changes in SpO_2 (blue) and heart rates (red) since birth, where a first desaturation occurred at 19:44 on his 3.73 days of life and first bradycardia at 20:27 can be observed. Three hours later, the neonatologist decided

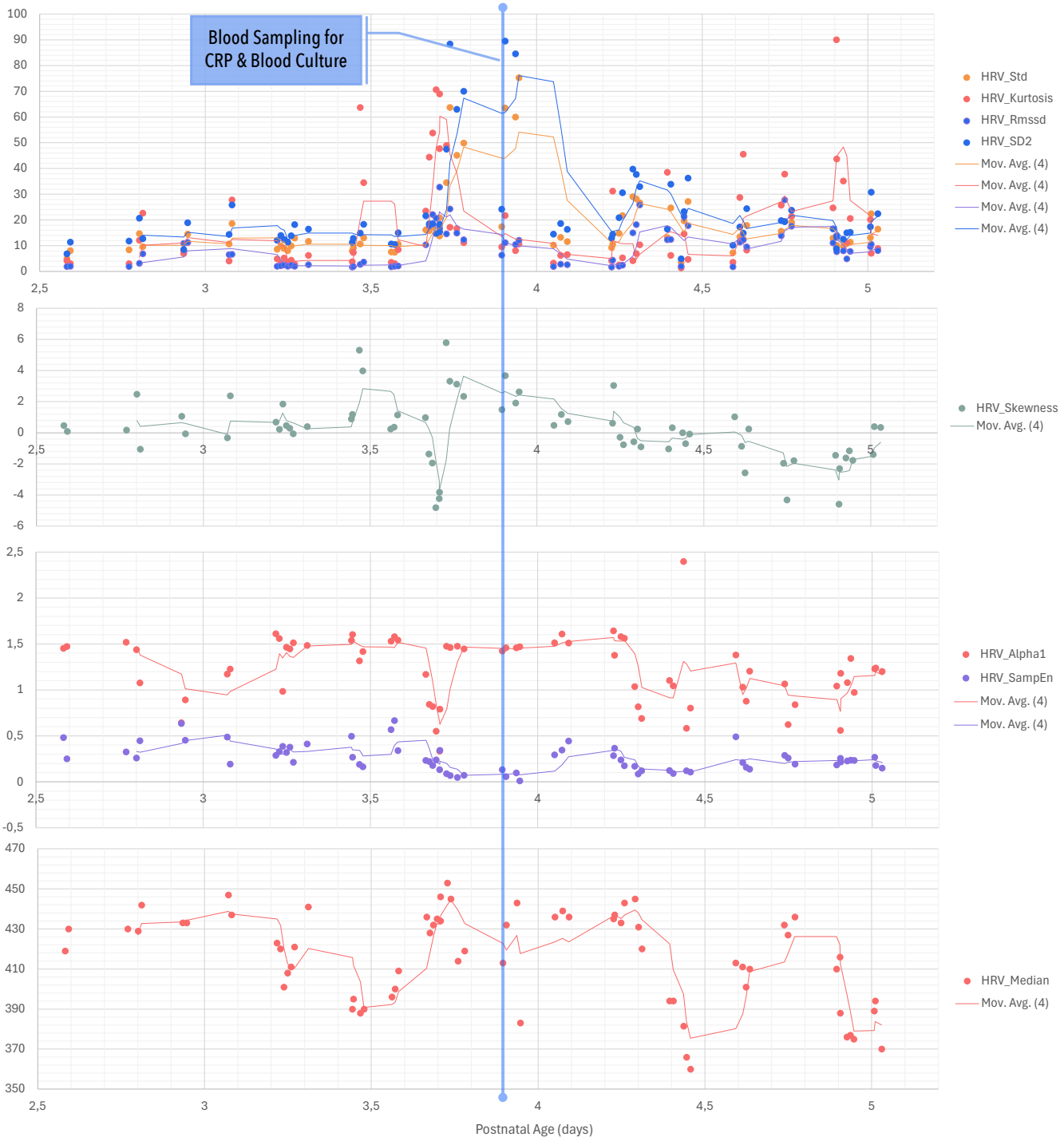

Figure 6.9: A clinical example of heart rate variability changes before clinical suspicion and confirmation of a sepsis event. The curves in solid lines are moving averages of 4 corresponding samples. The heart rate variability values are the results calculated in pseudo-real-time using the proposed on-the-edge CDSS and saved in **TEMP** folder.

to take a first blood sampling used for CRP and blood cultures for this newborn at 23:30 on the same day. And in another two hours, the first antibiotics were administrated for this newborn in the early morning of the next day.

The deployed on-the-edge CDSS keeps running and generates results of the quasi-real-time physiological data processing that are stored in the **OUT** mount point and the **TEMP** folder. By visualizing the heart rate variability parameters derived every 15 minutes of the same newborn, as shown in Figure 6.9, we could observe that this event was associated with an advanced increase in Std, Kurtosis, Rmssd, SD2 and Median; an advanced decrease in Skewness, $\alpha 1$ and Sample Entropy. Other HRV features not shown here had coherent variations. The neonatologists involved are well-trained and experienced, and they often intervene in suspected newborns before obtaining positive blood culture results for proven infections. However, these progressive changes in HRV appeared approximately 3 to 7 hours before blood collection (clinical suspicion) and before positive coagulase-negative staphylococcal blood cultures (clinical confirmation), suggesting that HRV may provide an earlier indication of sepsis prior to any clinical action are taken.

It is worth mentioning that all the results are produced in a quasi-real-time manner, suggesting the feasibility and meaningfulness of both the proposed on-the-edge system and the intermediate clinical results derived from the system, which is quite inspiring.

6.5 Discussion

To our knowledge, this is the first description of a multi-source, on-the-edge CDSS deployed in a NICU scenario. The quantitative technical performance evaluation of this prototype demonstrates the feasibility of the system's deployment in real-time hospital environments. This VM-based architecture highlights the features of usability, adaptability, and scalability, as well as its ability to be implemented effectively. Furthermore, the system's flexibility is enhanced by the ability to easily generalize the pipeline through simple model replacement using the ".pkl" format for inference. The integration of our previously published signal processing chain and ML models [14, 15] can be performed in this way.

In this implementation, we took advantage of existing instantaneous heart rate data from bedside monitors for subsequent HRV analysis. In further studies, we might also integrate our real-time signal processing methods into the feature-extraction phase. However, translating these algorithms into routine clinical practice demands close collaboration with medical device manufacturers to facilitate real-time execution on the bedside monitors, with rapid processing of raw physiological signals. Recent interactions between *LTSI - INSERM U1099*, *CHU Rennes* and Philips are going in this sense, targeting for implementation of research-level algorithms into the bedside monitors, through a specific application programming interface (API). On the other hand, when integrated into vendor-neutral systems, real-time signal extraction from patient monitors is required,

and the model must be robust to variations in the format of raw signals from different vendors, adding an additional interoperability step that might be particularly time-consuming. These challenges can pose significant barriers to large-scale deployment and may result in a performance decrease.

For this pilot study, hence, we leveraged the numerical data from Philips monitors to establish a proof-of-concept of a CDSS that incorporates a quasi-real-time processing engine and ML in inference mode for improving medical care in NICU settings. In fact, a recent study suggests that using numerical data of vital signs can achieve performance nearly equivalent to that of raw physiological signals in early detection of neonatal sepsis, with advantages in terms of storage consumption and computational efficiency, and without the need to develop specialized signal processing algorithms [16]. Nevertheless, our system is capable of obtaining raw ECG waveforms at the **IN** mount point by a frequency of every 1 hour. This transmission frequency can be increased by our Philips partners as demanded. In future development, we may consider including the high-resolution raw signals (500 Hz of sampling frequency) and applying our robust real-time QRS detector [17] on the signals, as described above. We might also integrate real-time respiratory signal processing, as performed in other works from our team.

One major limitation of the proposed system is related to the inference model embedded in the current version, which was trained on a different database (CARESS-Premi) without fine-tuning. Thus, its generalization and performance in estimating hyperbilirubinemia are limited. Further evaluation of the model's generalization and adaptation to this particular application is necessary, and prospective evaluations on clinical performance are thus warranted (protocol currently running). The important aspect to keep in mind here is that the model can be very easily modified through an updated PICKLE file, without altering the architecture of the proposed system.

Additionally, due to regulatory aspects and limited time, we did not finalize the last part of this architecture, which is to return the results generated by the system to the HIS. This full implementation would enable caregivers to directly access the results in real-time via the existing infrastructure in the hospital, potentially improving intervention strategies. Furthermore, the inclusion of alert and intervention mechanisms, such as advising phototherapy when estimated bilirubin levels exceed certain thresholds, would complete the close-loop system from **OUT** to the Hospital Information System (HIS), further enhancing clinical decision-making.

6.6 Conclusion

This chapter constructs an infrastructure in the NICU of a pilot clinical center (*CHU Rennes*) that can collect, store, exploit, analyze and present high-resolution physiological data from bedside monitors and the EHR system. The proposed system consisting of data transmission, pseudonymization, data fusion, processing and inference was deployed at the University Hospital of Rennes in

Jan 2023. Despite some limitations of the bilirubin estimation ML inference model itself, which was the first use case embedded into the on-the-edge system, a quantitative assessment of the technical performance of the system in terms of stability and resource consumption was achieved and a robust and satisfactory system configuration was obtained. A specific clinical case report based on the intermediate results derived from the system in a quasi-real-time manner provides inspiring implications on the system's usefulness and feasibility. The preliminary results greatly boost the confidence for future optimization and generalization of such infrastructure in a real NICU context. To our knowledge, this is the first description of a multi-source, on-the-edge CDSS deployed in a NICU scenario for patient-specific early detection of high-risk events. This proof-of-concept is a solid first step to initiate concrete on-the-edge clinical applications in the proposed platform, which can accommodate AI methods for patient-specific early detection of high-risk events by exploiting the dynamical properties of multivariate and multi-source longitudinal health data.

BIBLIOGRAPHY

[1] A. Rao and J. Palma, "Clinical decision support in the neonatal ICU," Seminars in Fetal and Neonatal Medicine, vol. 27, no. 5, p. 101332, 2022.

[2] N. Bressan, A. James, and C. McGregor, "Trends and opportunities for integrated real time neonatal clinical decision support," in Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, 2012, pp. 687–690.

[3] A. R. Spitzer, D. Ellsbury, and R. H. Clark, "The Pediatrix BabySteps® data warehouse — a unique national resource for improving outcomes for neonates," The Indian Journal of Pediatrics, vol. 82, no. 1, pp. 71–79, 2015.

[4] E. Meyeroff and P. Tremoulet, "Etiometry's T3 heuristic evaluation," Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, vol. 10, no. 1, pp. 37–41, 2021.

[5] A. Asfari, "Artificial intelligence role and clinical decision support system extubation readiness trail and Etiometry scoring system," Biomedical Journal of Scientific & Technical Research, vol. 35, no. 1, pp. 27 291–27 293, 2021.

[6] H. Singh, R. Mallaiah, G. Yadav, N. Verma, A. Sawhney, and S. K. Brahmachari, "iCHRCloud: web & mobile based child health imprints for smart healthcare," Journal of Medical Systems, vol. 42, no. 1, p. 14, 2018.

[7] H. Singh, R. Kaur, A. Gangadharan, A. K. Pandey, A. Manur, Y. Sun, S. Saluja, S. Gupta, J. P. Palma, and P. Kumar, "Neo-bedside monitoring device for integrated neonatal intensive care unit (iNICU)," IEEE Access, vol. 7, pp. 7803–7813, 2018.

[8] D. R. Kommers, R. Joshi, C. van Pul, L. Feijs, S. Bambang Oetomo, and P. Andriessen, "Changes in autonomic regulation due to kangaroo care remain unaffected by using a swaddling device," Acta Paediatrica, vol. 108, no. 2, pp. 258–265, 2019.

[9] M. Wilken, D. Hüske-Kraus, and R. Röhrig, "Alarm fatigue: using alarm data from a patient data monitoring system on an intensive care unit to improve the alarm management," in German Medical Data Sciences: Shaping Change–Creative Solutions for Innovative Medicine. IOS Press, 2019, pp. 273–281.

[10] R. Joshi, D. Kommers, L. Oosterwijk, L. Feijs, C. van Pul, and P. Andriessen, "Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics, and ECG-derived estimates of infant motion," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 3, pp. 681–692, 2020.

[11] Z. Peng, G. Varisco, X. Long, R.-H. Liang, D. Kommers, W. Cottaar, P. Andriessen, and C. van Pul, "A continuous late-onset sepsis prediction algorithm for preterm infants using multi-channel physiological signals from a patient monitor," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 1, pp. 550–561, 2023.

[12] M. Hartmann, U. S. Hashmi, and A. Imran, "Edge computing in smart health care systems: Review, challenges, and research directions," Transactions on Emerging Telecommunications Technologies, vol. 33, no. 3, p. e3710, 2022.

[13] J. Kabachinski, "What is health level 7?" Biomedical Instrumentation & Technology, vol. 40, no. 5, pp. 375–379, 2006.

[14] N. Montazeri Ghahjaverestan, M. B. Shamsollahi, D. Ge, A. Beuchée, and A. I. Hernández, "Apnea bradycardia detection based on new coupled hidden semi Markov model," Med Biol Eng Comput, vol. 59, no. 1, pp. 1–11, 2021.

[15] M. Chen, A. Beuchée, F. Tudoret, and A. I. Hernández, "Non-invasive total serum bilirubin estimation in preterm infants with modified mixed-effects random forest," in 2024 Computing in Cardiology (CinC). IEEE, In Press.

[16] M. Yang, Z. Peng, C. van Pul, P. Andriessen, K. Dong, D. Silvertand, J. Li, C. Liu, and X. Long, "Continuous prediction and clinical alarm management of late-onset sepsis in preterm infants using vital signs from a patient monitor," Computer Methods and Programs in Biomedicine, p. 108335, 2024.

[17] M. Doyen, D. Ge, A. Beuchée, G. Carrault, and A. I. Hernández, "Robust, real-time generic detector based on a multi-feature probabilistic method," PLoS ONE, vol. 14, no. 10, p. e0223785, 2019.

Conclusions and perspectives

6.7 Conclusions

Neonatal intensive care units (NICU) are specialized services in hospitals that provide intensive care to newborn babies who are born prematurely, critically ill, or require close monitoring and treatment. These units are equipped with advanced life-support equipment and staffed by experienced neonatologists, nurses, and other healthcare professionals. NICU are a real hub of activity, generating massive amounts of data on a daily basis. From electronic health records to vital sign monitors and their alarms, imaging studies, and laboratory results, the sheer volume of information can be overwhelming for healthcare professionals. These data are often fragmented across different systems and formats, making it challenging for clinicians to access, analyze, and integrate the information they need. The result is a significant burden on healthcare professionals' time and attention, as they struggle to make sense of this complex data landscape. With so much at stake—from patient outcomes to quality improvement initiatives—effective management of these massive amounts of data is crucial. This data pressure demands innovative solutions to manage and analyze this information. Machine learning (ML) and on-the-edge processing are two powerful technologies that can help healthcare professionals cope with these massive amounts of data and be able to personalize monitoring and treatment.

In this context, this dissertation has explored the proposal, the development and the application of advanced data processing techniques and interpretable machine learning models to tackle two critical challenges in neonatal care: Neonatal hyperbilirubinemia (NHB) and Late-onset sepsis (LOS) in preterm infants. By integrating real-life monitoring data with non-invasive approaches, we aimed to improve early diagnosis and facilitate clinical decision-making, fostering more personalized and effective neonatal care in the NICU.

As a main building block, the first contribution of this work was the integration and improvement of an automated cardiac signal processing pipeline tailored for NICU environments (Chapter 2) based on the previous work of our team *SEPIA* of *LTSI - INSERM U1099*. The proposed pipeline processes real-world ECG data from preterm infants, significantly reducing manual labor while ensuring high-quality, clinically relevant outputs. It includes several critical steps: ECG signal quality evaluation, QRS complex detection, RR interval extraction, RR correction, stationarity analysis, and heart rate variability analysis. This end-to-end solution is designed to handle noisy and artifact-prone signals common in NICU settings, particularly for preterm infants with low yet

complex cardiovascular variability. By integrating these components, the pipeline enables more reliable long-term monitoring and produces utile parameters for further analysis and machine learning model development, while greatly reducing the labor and time included. This pipeline served as a fundamental tool used throughout the dissertation.

For the management of Neonatal hyperbilirubinemia (NHB), we investigated two important aspects: A) model-based characterization of Total serum bilirubin (TSB) dynamics, and B) knowledge-based non-invasive estimation of TSB using mixed-effects machine learning models. Together, these studies aimed to offer a comprehensive approach to tackling the clinical challenges of hyperbilirubinemia in preterm infants and to pave the way for more effective and patient-specific clinical interventions.

A) Model-based characterization of TSB dynamics (Chapter 3)
Another contribution of this work was the proposal and validation of a patient-specific exponential decay model to characterize the natural and long-term dynamics of TSB concentrations in preterm infants born between 24 and 32 weeks of gestation. Through personalized model parameter fitting, we obtained 72 models with patient-specific parameters optimized by minimizing the error between measured TSB and model output using an adaptive robust least-squared method. The proposed model demonstrated its effectiveness and capability to closely track observed TSB levels during extended neonatal periods, with an RMSE ranging from 1.20 to 40.25 µmol/L, with a median [IQR] of 8.74 [4.89; 14.25] µmol/L. Furthermore, when the bilirubin evolutionary trend of a given patient diverges from the expected decay pattern, as indicated by an increased RMSE, it might suggest the occurrence of high-risk clinical events such as necrotizing enterocolitis and elevated C-reactive protein levels. This association indicates that the model's capabilities extend beyond mere descriptive analytics and may serve as a new digital tool for the early detection of relevant comorbidities. This contribution has been published in an international journal.

B) Knowledge-based TSB estimation using mixed-effects ML models (Chapter 4)
We explored non-invasive approaches for estimating TSB levels in preterm infants with/without hyperbilirubinemia born at $24^{2/7}$ to $31^{6/7}$ gestational weeks. The main methodological problems in this context were related to the correct integration of the longitudinal aspect of these data into the ML process and the integration of explicit models of the underlying physiology. We proposed and compared different hybrid machine learning estimators with the incorporation of mixed effects and physiological knowledge representation. When compared to a standard Random Forest (RF), the proposed modified Mixed-Effects Random Forest (MERF) models greatly improved the estimation agreements and reduced the proportional bias thanks to the explicit integration of meaningful physiological knowledge. Although the proposed hybrid models require patient-specific historical data for their initialization and the model performance is still far from being compatible with an actual clinical application, they show clinical potential in the NICU where longitudinal clinical data are commonly seen. This contribution has been published in an international conference.

Another major challenge in the NICU is the early detection of neonatal Late-onset sepsis (LOS) (Chapter 5). One first contribution to this axis was the formalization of the clinical timeline, with time constants estimated from the literature, clinical expertise, or directly from our CARESS-Premi dataset. This formalization eases the representation of the causal effects that are involved during the LOS decision-making and allows us to propose an original approach in this field. Indeed, in this work, we evaluated the efficiency of sepsis risk detection that is made without taking clinical suspicion (i.e., before treatment and intervention begin) into account when training models, which may be a lack of investigation in the literature.

A second contribution to this axis was the constitution of a formally annotated real-life longitudinal database of multi-parametric signal monitoring in NICU for LOS early detection. We retrieved continuous and longitudinal monitoring ECG signals of around 450 preterm infants and segmented them into 6-hour blocks. The proposed signal processing pipeline was performed for all data segments to derive interested HRV features. A labeling strategy was designed to generate pseudo labels for each segment that indicate the patient status (sepsis, non-sepsis or uncertain). This database might be used in further research in our team, in particular for the joint exploitation of ECG and respiratory signals.

A third contribution in this axis was the proposal and evaluation of the above-mentioned database of a range of machine learning models, including Logistic Regression (LR), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Multilayer perceptron (MLP) as well as an original shallow Convolutional neural networks (Shallow CNN). According to the degree of time independence, the proposed detectors could be categorized as instantaneous detectors (time-independent) and time-dependent detectors. The highest AUROC and AUPRC among instantaneous detectors was achieved by an RF classifier, which were 0.727 ± 0.060 and 0.352 ± 0.099, respectively; while the shallow CNNs obtained the best performance in a short-term time-dependent manner: 0.737 ± 0.027 of AUROC (used three successive data segments) and 0.749 ± 0.045 of AUROC (used six successive data segments). It is admitted that our results were not comparable to the high-performance levels shown in some papers in the literature (AUROC > 0.80). However, and more importantly, we claim that it is impossible to make direct comparisons between the studies since there are huge heterogeneity and wide variability in many aspects such as the definitions of neonatal sepsis, the annotation strategies, etc. We also criticized the way in which most of these works from the literature were built and evaluated, and we proposed a thorough discussion on the challenges related to a correct performance evaluation in this context. Two particular limitations are highlighted: *i)* the lack of a precise, standardized definition of neonatal LOS and *ii)* the very high potential of underestimation of false positives in these works. We consider that supervised ML model training in most of these publications suffers from significant sources of bias, which may explain their lack of clinical applicability and generalization, and thus a significant effort should be made in this field to foresee useful clinical applications in the future. This discussion, as well as the efforts performed in this work for the formal, time-sensitive annotation and processing of the

CARESS-Premi database, are for us an additional contribution to this dissertation.

The final contribution of this work was the proposal of a proof-of-concept CDSS that facilitates early detection, diagnosis and intervention in a clinical context (Chapter 6). We designed, implemented, deployed, and technically evaluated a CDSS that integrates quasi-real-time signal processing chains and ML models on the edge in a NICU setting. The proposed system consisting of data transmission, pseudonymization, data fusion, processing, and inference was deployed at *CHU Rennes* since Jan 2023. During the first six months of deployment, the service continuously received monitoring signal data from 138 neonates, processing live data and generating bilirubin level estimations at a temporal resolution of 15 minutes. A quantitative assessment of the technical performance of the system in terms of stability and resource consumption was achieved and a robust and satisfactory system configuration was obtained. In addition, a specific clinical case report on the predictive properties of HRV before the clinical suspicion of sepsis, based on intermediate results obtained by the system in a near real-time manner, provides inspiring implications on the practicality and feasibility of the system. To our knowledge, this is the first description of a multi-source, on-the-edge CDSS deployed in a NICU scenario for patient-specific early detection of high-risk events. This proof-of-concept is a solid first step to initiate concrete on-the-edge clinical applications by exploiting the dynamical properties of multivariate and multi-source longitudinal health data in the proposed platform. This contribution has been published in an international conference.

In conclusion, by leveraging physiological signal processing and machine learning techniques, this dissertation proposed different knowledge-based and model-based approaches for improving the diagnosis and management of critical neonatal conditions such as hyperbilirubinemia and late-onset sepsis. Moreover, an on-the-edge CDSS, as a proof-of-concept, was successfully deployed and preliminarily evaluated in a real-life NICU setting, demonstrating the effectiveness of the proposed pipeline and models working in real-time. Overall, the contributions in this dissertation show significant promise for optimizing NICU monitoring and improving early detection of high-risk events. These advancements could greatly reinforce the management and outcomes for preterm infants.

6.8 Perspectives

This dissertation has laid a solid foundation for improving neonatal care using data-driven methods, but it also opens avenues for further research and development.

On top of all, event detection remains both clinically complicated and methodologically challenging due to the lack of universally accepted definitions, annotations, and all difficulties discussed throughout the dissertation. If we zoom out to see the big picture, large efforts should be made to enhance the *Gold Standards* on event detection in the NICU, such as:

- Establish comprehensive and standardized clinical definitions.
- Develop collaborative frameworks that integrate expertise from multiple centers to create robust and consistent annotations.
- Enhance the interpretability of event detection algorithms, ensuring their alignment with clinical practices and patient outcomes.

Specific to this dissertation, there are some actionable aspects for improving the proposed model personalization in performance. By taking our non-invasive bilirubin estimation as a starting point, we could focus on several aspects to achieve this:

- Hybrid TSB estimation models:
  - Introduce different coefficients of random effects ($b_{1i}Z1 + b_{2i}g(Z2)$) to better capture inter-patient variability;
  - Refine the EM process in MERF learning to improve model robustness and parameter optimization.
- Personalized the non-linear term in random effects:
  Transition from a global function $g(\cdot)$ to patient-specific functions $g_i(\cdot)$, allowing for greater adaptability and precision in predictions.
- Extensive data collection:
  Broaden the dataset by including diverse patient populations and expanding to multi-center studies, ensuring the models' validity and reliability across various clinical settings.

Last but not the least, the on-the-edge clinical decision support systems developed in this dissertation provides a promising framework for real-time neonatal care. To maximize its potential, future work should focus on the completion and enrichment of the on-the-edge CDSS:

- Closing the loop of the proposed system architecture:
  Integrate/push the CDSS outputs with Hospital Information System (HIS) to enable seamless feedback and clinical workflow integration.
- Integrating new models in inference mode:
  Deploy advanced models, such as hybrid mathematical and machine learning models, for a wider range of clinical applications, including sepsis detection and bilirubin monitoring. This helps to share the system infrastructure with team members and collaborators to facilitate model deployment and usage across different NICU.
- Multi-center prospective clinical evaluation:
  - Address the challenges of unbiased validation in real-world settings by conducting prospective clinical trials across multiple centers;
  - Align these trials with emerging regulatory standards, such as the EU AI Act, to ensure compliance and scalability for clinical adoption, for example, leverage initiatives like the ANR EXPERT project to validate the CDSS on a larger scale.

These perspectives outlined above highlight both the scientific and translational potential of this dissertation. By addressing the challenges and advancing the proposed directions, we believe that this work can contribute notably to the field of neonatal care, fostering the development of more personalized, effective, and clinically meaningful interventions for preterm infants in NICUs.

List of Publications

International Journal Papers

[1] **M. Chen**, A I. Hernández, "Towards an explainable model for sepsis detection based on sensitivity analysis," *IRBM*, vol. 43, no. 1, pp. 75–86, 2022, doi:10.1016/j.irbm.2021.05.006.

[2] **M. Chen**, A. Beuchée, E. Levine, L. Storme, G. Gascoin, A I. Hernández, "Model-based characterization of total serum bilirubin dynamics in preterm infants," *Pediatric Research*. Published online November 7, 2024. doi:10.1038/s41390-024-03644-z.

International Conferences

[1] **M. Chen**, A. Beuchée, F. Tudoret, A. Coursin, P. Ho, and A I. Hernández, "Deployment of an on-the-edge clinical decision support system in neonatal intensive care units," in *2023 Computing in Cardiology (CinC)*, vol. 50. IEEE, 2023, pp. 1-4, doi:10.22489/CinC.2023.061.

[2] **M. Chen**, A. Beuchée, F. Tudoret, and A I. Hernández, "Non-invasive total serum bilirubin estimation in preterm infants with modified mixed-effects random forest," in *2024 Computing in Cardiology (CinC)*, vol. 51. IEEE, 2024, pp. 1-4, doi:10.22489/CinC.2024.120.

List of Figures

List of Tables

Appendices

Table A.1: General characteristics of the included population and the excluded population (patients with fewer than four bilirubin measurements).

	Characteristics	Number of Records	Overall Population N=373	Included Patients N=288	Excluded Patients N=85	P-value
Initial conditions	Multiple pregnancy, n (%)	104 (27.9%)	87 (30.2%)	17 (20.0%)	0.088	
	Hypertension in pregnancy, n (%)	373	89 (23.9%)	54 (18.8%)	35 (41.2%)	$<0.001^{\star}$
	Preterm labor, n (%)	373	225 (60.3%)	192 (66.7%)	33 (38.8%)	$<0.001^{\star}$
	Chorioamnionitis, n (%)	373	22 (5.90%)	19 (6.60%)	3 (3.53%)	0.428
	Corticosteroids, n (%)	373	349 (93.6%)	268 (93.1%)	81 (95.3%)	0.626
	Delivery route, n (%)	373				0.018
	Vaginal delivery		144 (38.6%)	121 (42.0%)	23 (27.1%)	
	C-section		229 (61.4%)	167 (58.0%)	62 (72.9%)	
	GA at birth (weeks), mean (SD); median	373	28.5 (1.82); 28.6	28.1 (1.74); 28.1	29.9 (1.42); 28.9	$< 0.001^{\star}$
	Birth weight (g), mean (SD); median	373	1133 (316); 1110	1082 (295); 1050	1306 (327); 1275	$< 0.001^{\star}$
	Birth weight Z-score†, mean (SD); median	373	−0.05 (0.79); 0.02	−0.03 (0.80); 0.05	−0.11 (0.75); −0.14	0.211
	Gender (male), n (%)	373	199 (53.4%)	160 (55.6%)	39 (45.9%)	0.148
	Apgar (1 min score), median [IQR]	370	6.00 [3.00; 8.00]	6.00 [3.00; 8.00]	6.00 [4.00; 8.00]	0.963
	Intubation at birth, n (%)	373	73 (19.6%)	62 (21.5%)	11 (12.9%)	0.110
	PDA on PNA = 4 days, n (%)	373	160 (42.9%)	139 (48.3%)	21 (24.7%)	$< 0.001^{\star}$

Table A.1 – Continued

	Characteristics	Number of Records	Overall Population *N*=373	Included Patients *N*=288	Excluded Patients *N*=85	*P*-value
Outcomes	Neurologic impairment, n (%)	373	84 (22.5%)	76 (26.4%)	8 (9.41%)	0.001*
	Respiratory support stopped before 34 PMA weeks, n (%)	373	119 (31.9%)	77 (26.7%)	42 (49.4%)	< 0.001*
	Death, n (%)	373	17 (4.56%)	16 (5.56%)	1 (1.18%)	0.160
	PNA at death (days), median [IQR]	17	29.5 [19.1; 66.0]	35.0 [19.5; 69.6]	15.4 [15.4; 15.4]	0.353
	PMA at death (weeks), median [IQR]	17	32.4 [28.8; 36.0]	33.2 [28.7; 36.6]	30.6 [30.6; 30.6]	0.941
	Interruption of follow-up, n (%)	373	2 (0.54%)	2 (0.69%)	0 (0.00%)	1.000
	Length of follow-up (days), mean (SD); median	372	26.1 (16.8); 23.4	30.0 (16.1); 27.2	13.2 (10.7); 8.42	< 0.001*
	Phototherapy, n (%)	373	344 (92.2%)	273 (94.8%)	71 (83.5%)	0.001*

* Statistical significance between included and excluded populations ($p < 0.0025$ after Bonferroni correction).

† Z-scored birth weight based on gestational age according to Fenton's 2013 preterm growth chart [21].

Features are divided into initial conditions and outcomes according to the time of acquisition before or after PNA = 4 days.

The statistics of patient characteristics were reported as **counts (percentage)** for categorical variables and as **mean (SD); median** or **median [IQR]** for continuous variables, as appropriate. Categorical variables were compared using the chi-square test and continuous variables using the Mann-Whitney U test, as appropriate.

GA: gestational age. *PDA*: patent ductus arteriosus. *PNA*: postnatal age. *PMA*: postmenstrual age. *SD*: standard deviation. *IQR*: interquartile range.

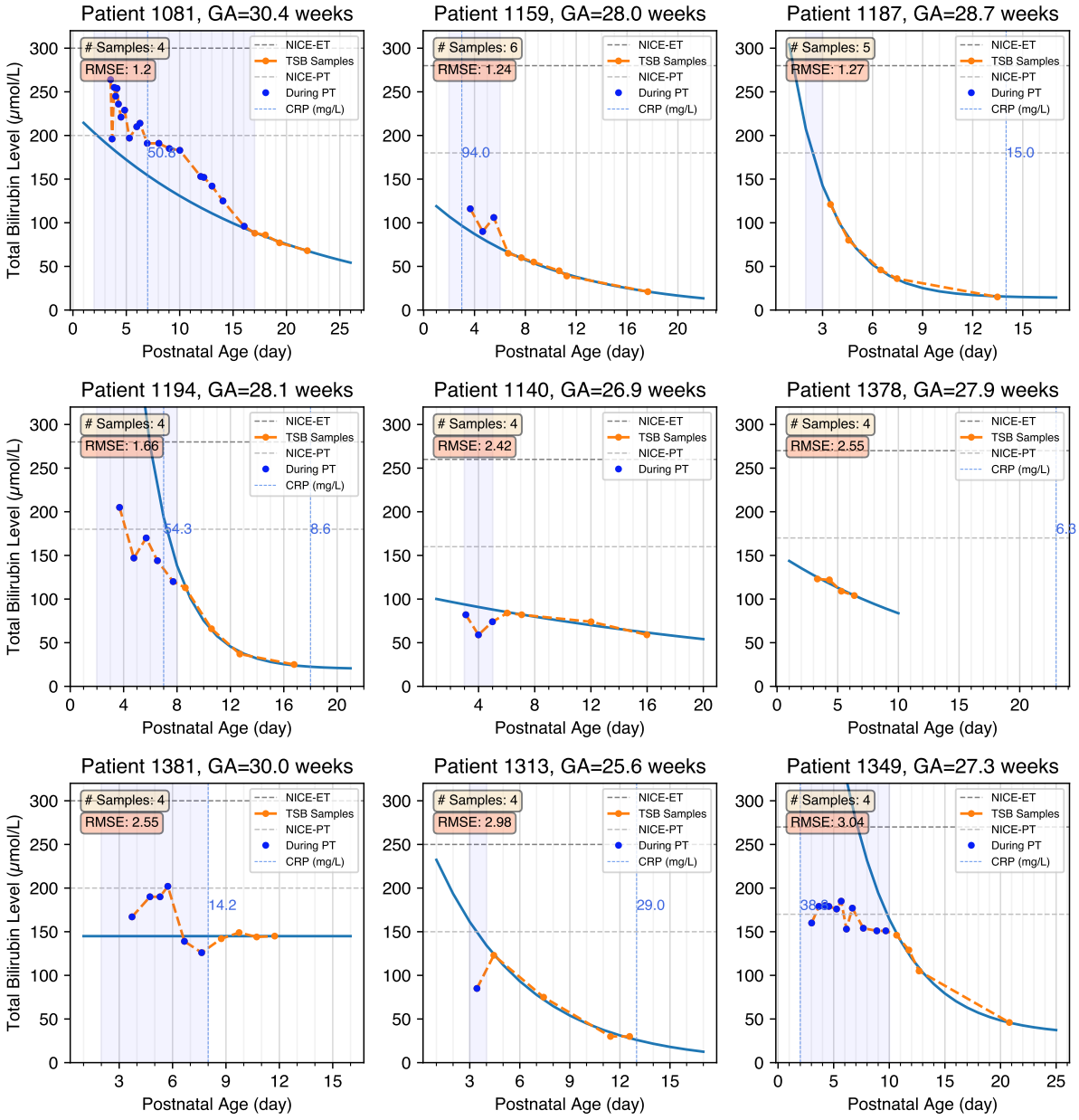

Figure A.1: Patient-specific total serum bilirubin exponential decay models on the PNA-TSB plane, ordered by increasing fitting errors (RMSE from 1.20 to 3.04 μmol/L).
The x-axes are postnatal age (PNA) in days; the y-axes are total serum bilirubin (TSB) in μmol/L. Modeled exponential decay curves are plotted as blue solid curves and TSB measurements are orange scatter points. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

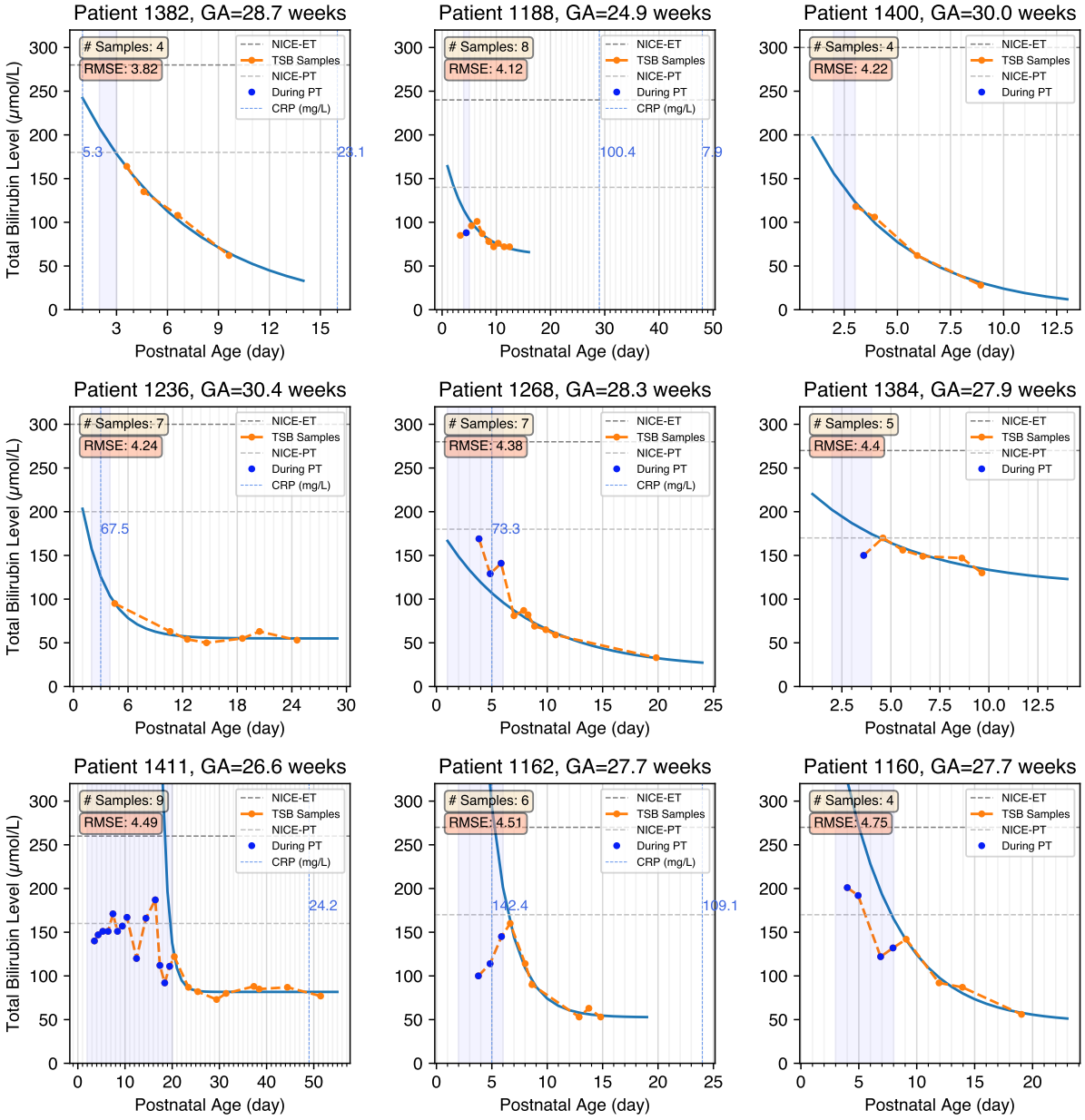

Figure A.2: Patient-specific total serum bilirubin exponential decay models on the PNA-TSB plane, ordered by increasing fitting errors (RMSE from 3.82 to 4.75 µmol/L).
The x-axes are postnatal age (PNA) in days; the y-axes are total serum bilirubin (TSB) in µmol/L. Modeled exponential decay curves are plotted as blue solid curves and TSB measurements are orange scatter points. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

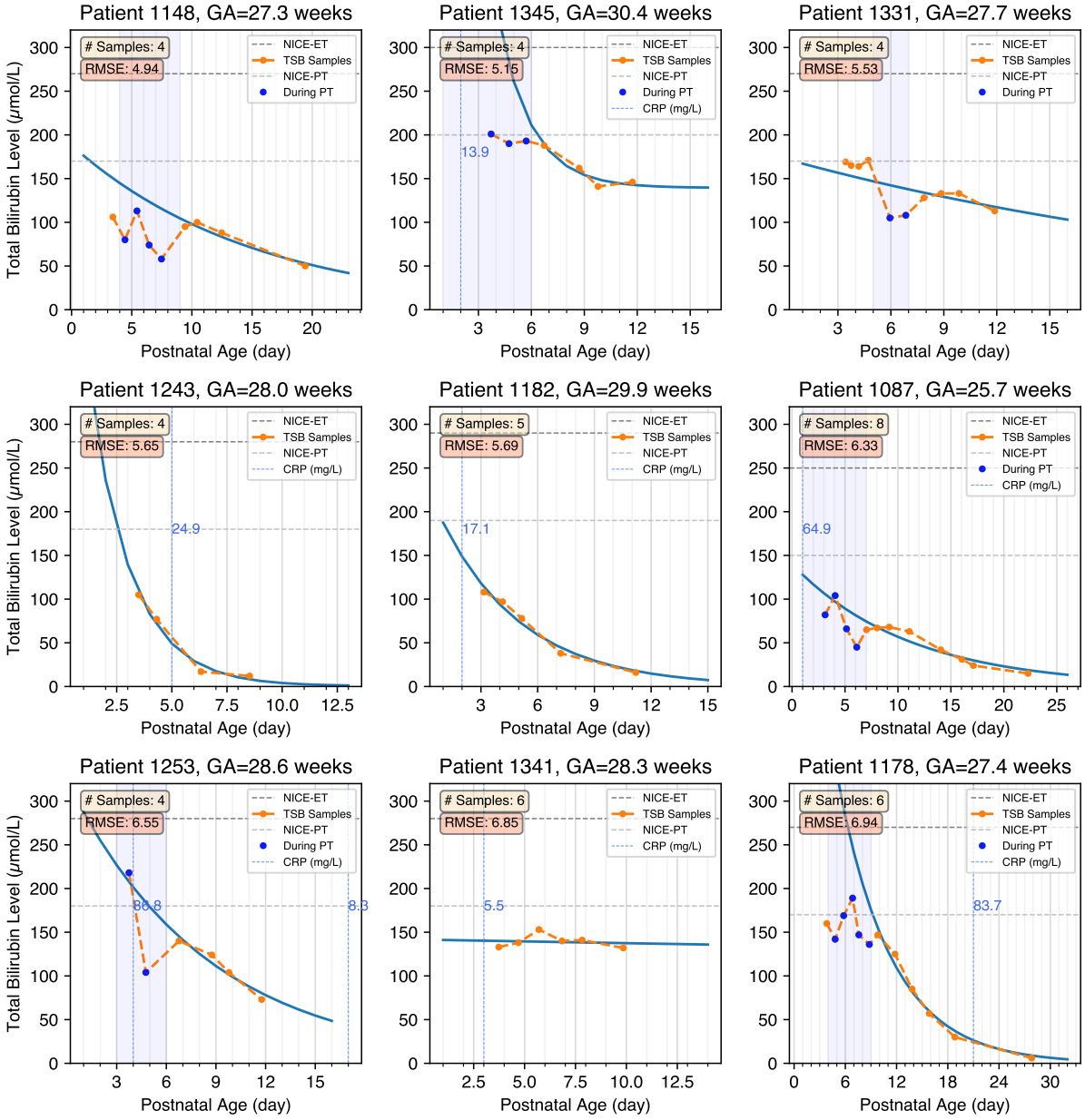

Figure A.3: Patient-specific total serum bilirubin exponential decay models on the PNA-TSB plane, ordered by increasing fitting errors (RMSE from 4.94 to 6.94 μmol/L).
The x-axes are postnatal age (PNA) in days; the y-axes are total serum bilirubin (TSB) in μmol/L. Modeled exponential decay curves are plotted as blue solid curves and TSB measurements are orange scatter points. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

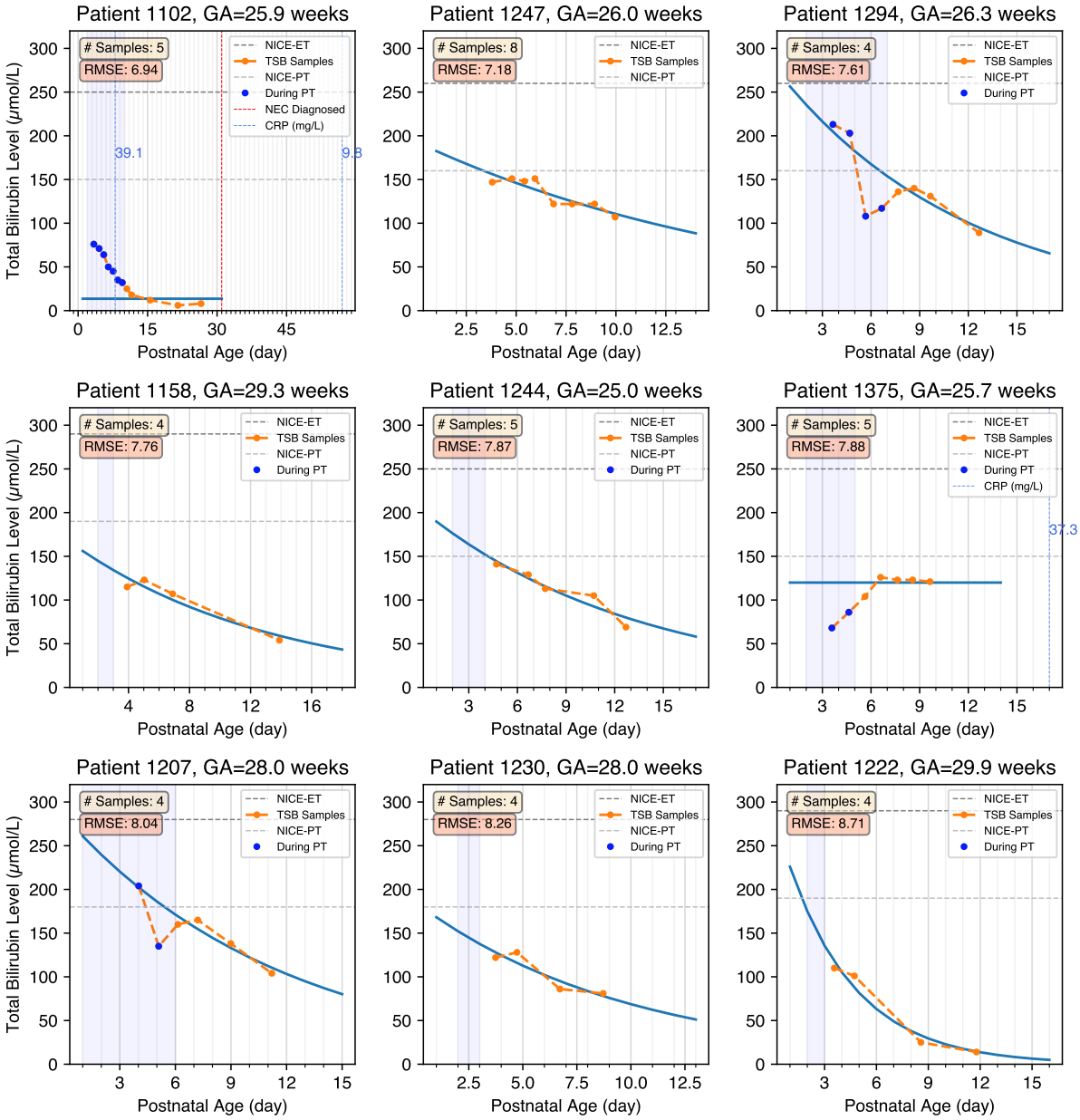

Figure A.4: Patient-specific total serum bilirubin exponential decay models on the PNA-TSB plane, ordered by increasing fitting errors (RMSE from 6.94 to 8.71 µmol/L).
The x-axes are postnatal age (PNA) in days; the y-axes are total serum bilirubin (TSB) in µmol/L. Modeled exponential decay curves are plotted as blue solid curves and TSB measurements are orange scatter points. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

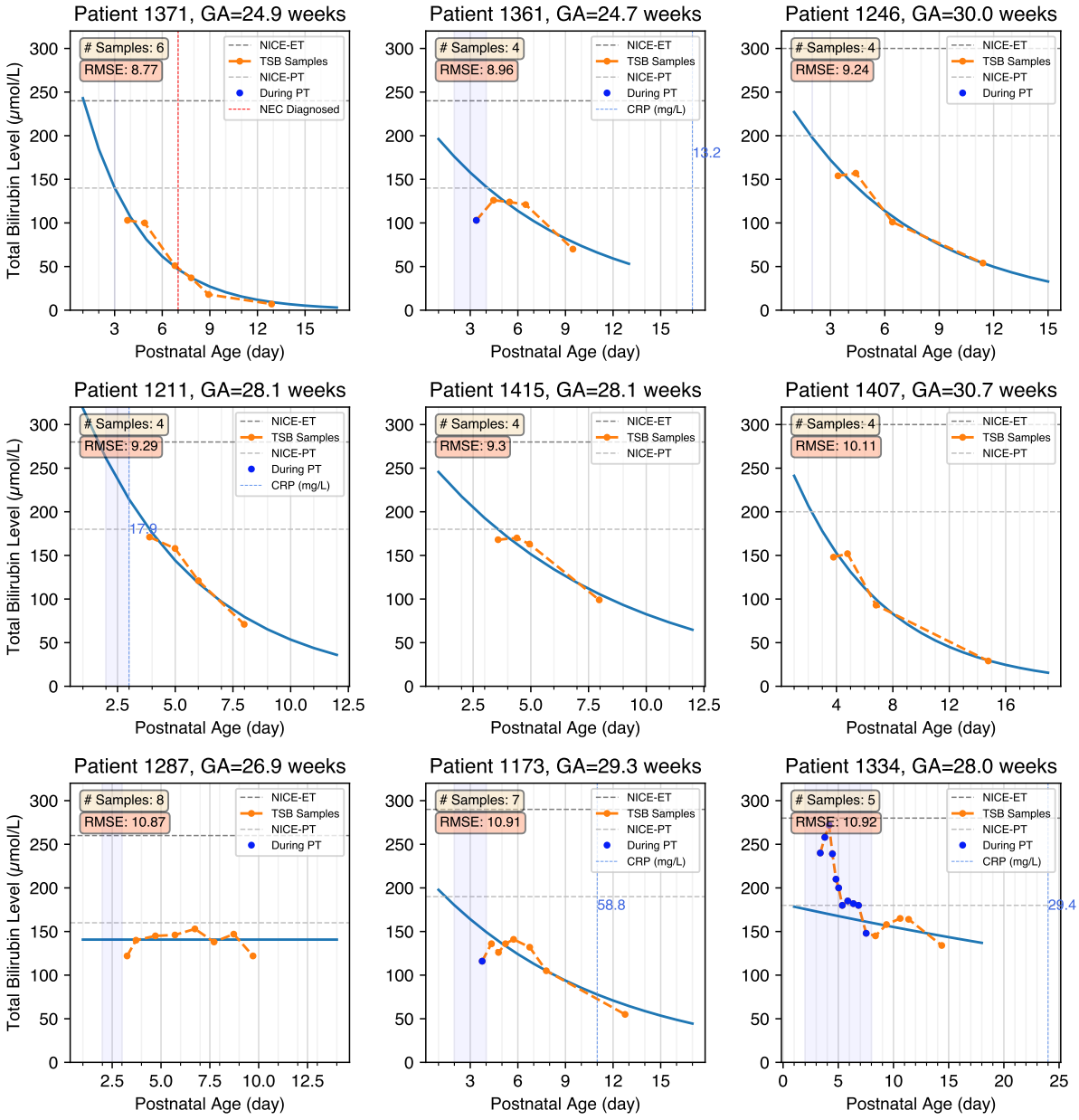

Figure A.5: Patient-specific total serum bilirubin exponential decay models on the PNA-TSB plane, ordered by increasing fitting errors (RMSE from 8.77 to 10.92 μmol/L).
The x-axes are postnatal age (PNA) in days; the y-axes are total serum bilirubin (TSB) in μmol/L. Modeled exponential decay curves are plotted as blue solid curves and TSB measurements are orange scatter points. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

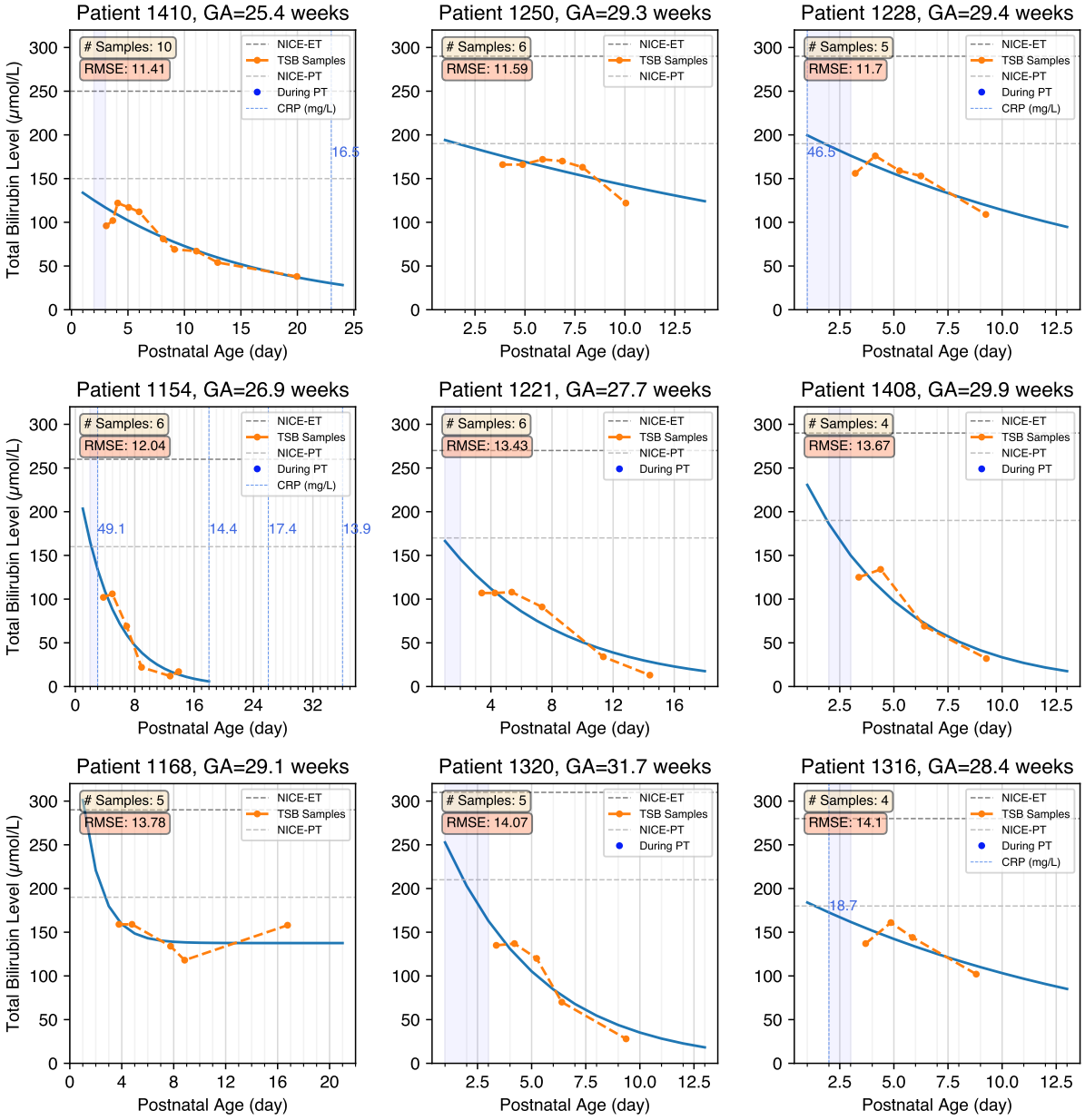

Figure A.6: Patient-specific total serum bilirubin exponential decay models on the PNA-TSB plane, ordered by increasing fitting errors (RMSE from 11.41 to 14.10 µmol/L).
The x-axes are postnatal age (PNA) in days; the y-axes are total serum bilirubin (TSB) in µmol/L. Modeled exponential decay curves are plotted as blue solid curves and TSB measurements are orange scatter points. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

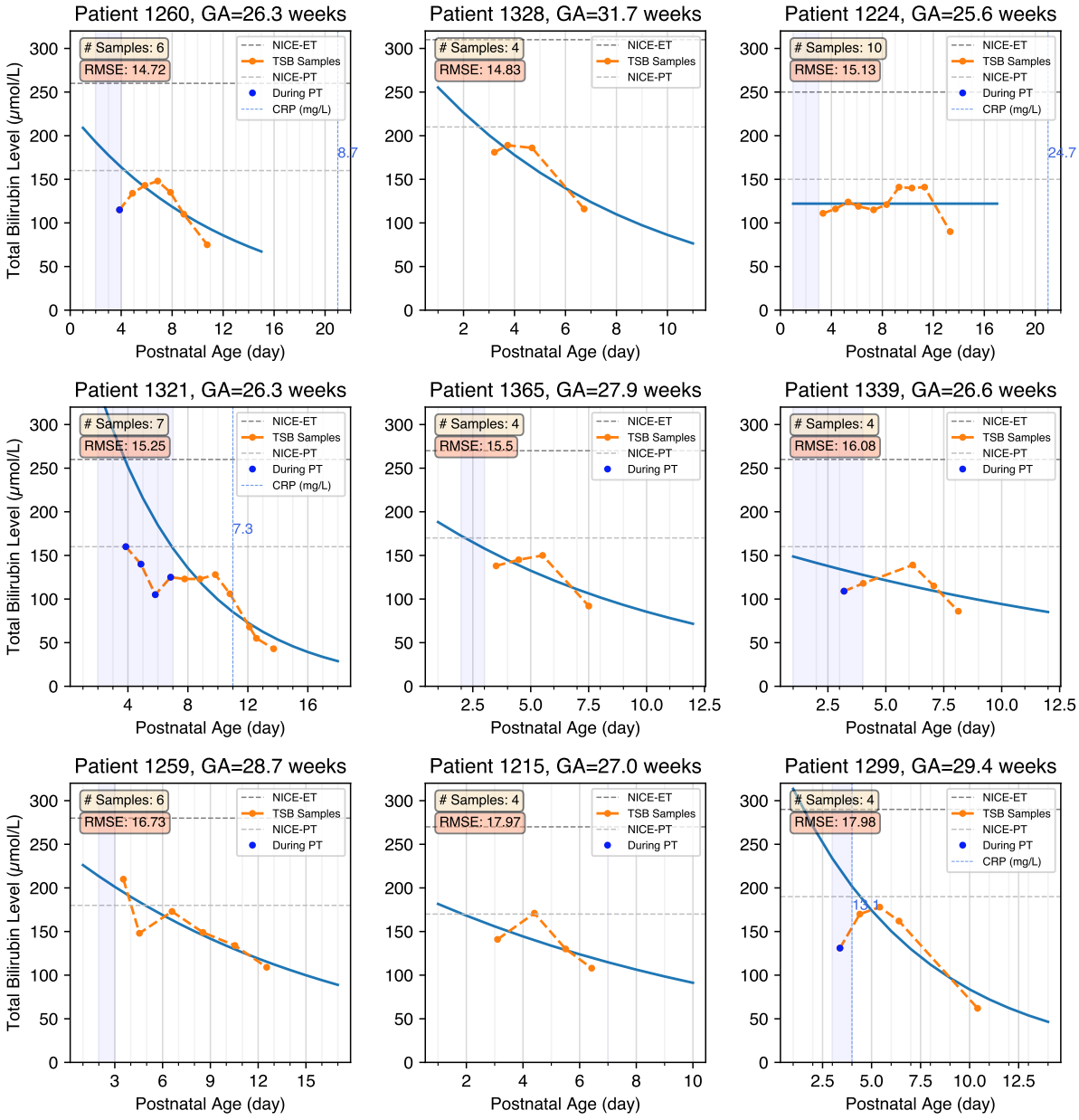

Figure A.7: Patient-specific total serum bilirubin exponential decay models on the PNA-TSB plane, ordered by increasing fitting errors (RMSE from 14.72 to 17.98 μmol/L).
The x-axes are postnatal age (PNA) in days; the y-axes are total serum bilirubin (TSB) in μmol/L. Modeled exponential decay curves are plotted as blue solid curves and TSB measurements are orange scatter points. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

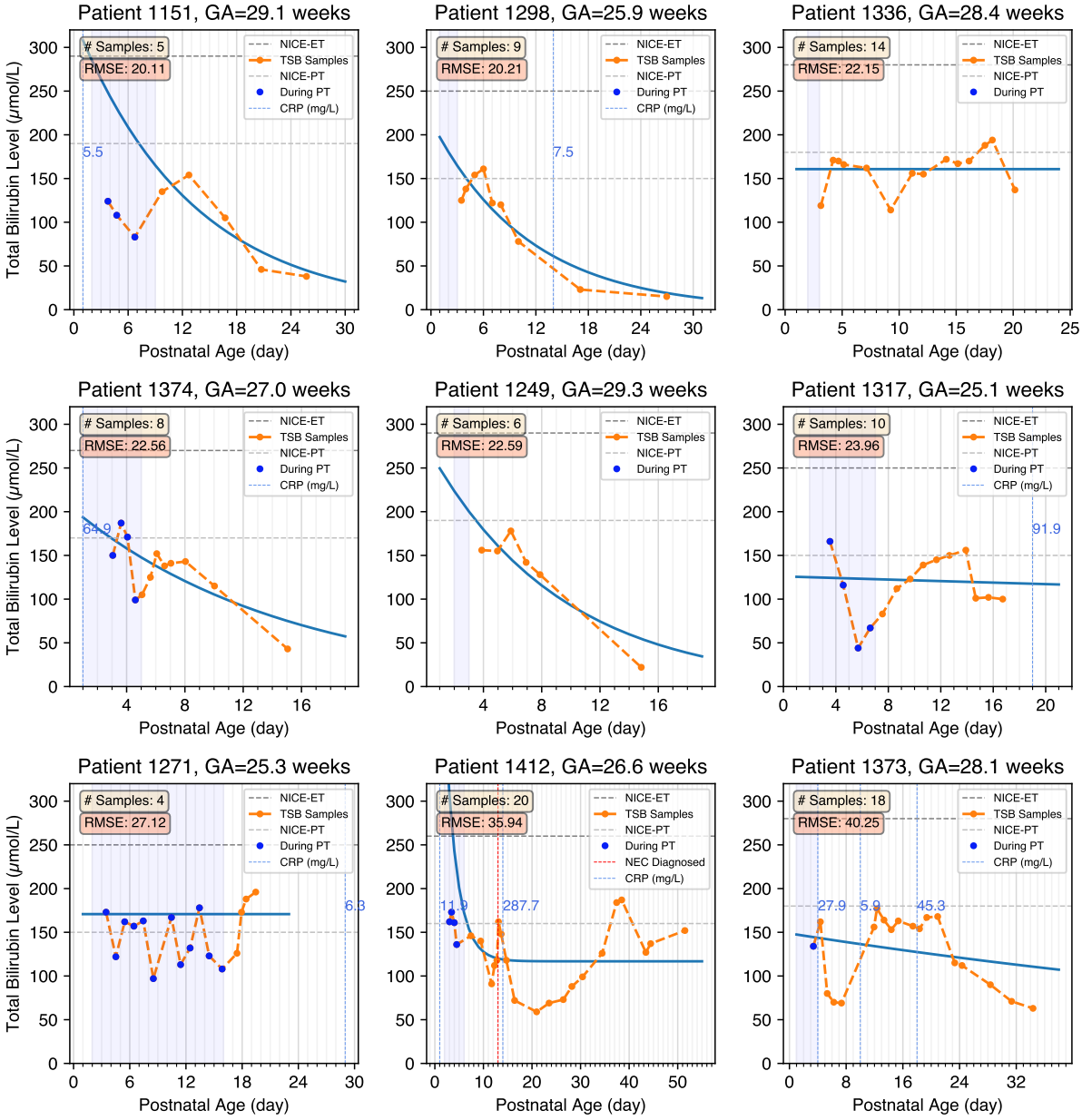

Figure A.8: Patient-specific total serum bilirubin exponential decay models on the PNA-TSB plane, ordered by increasing fitting errors (RMSE from 20.11 to 40.25 µmol/L).
The x-axes are postnatal age (PNA) in days; the y-axes are total serum bilirubin (TSB) in µmol/L. Modeled exponential decay curves are plotted as blue solid curves and TSB measurements are orange scatter points. Blue shades indicate phototherapy durations and TSB measurements performed during PT are marked in blue. Annotations in light blue next to blue dashed vertical lines are C-reactive protein (CRP) values measured on corresponding PNAs. Horizontal dashed lines are GA-specific thresholds for treatments according to NICE guidelines: PT (light grey) and ET (dark grey).
ET: exchange transfusion. *PT*: phototherapy. *NEC*: necrotizing enterocolitis.

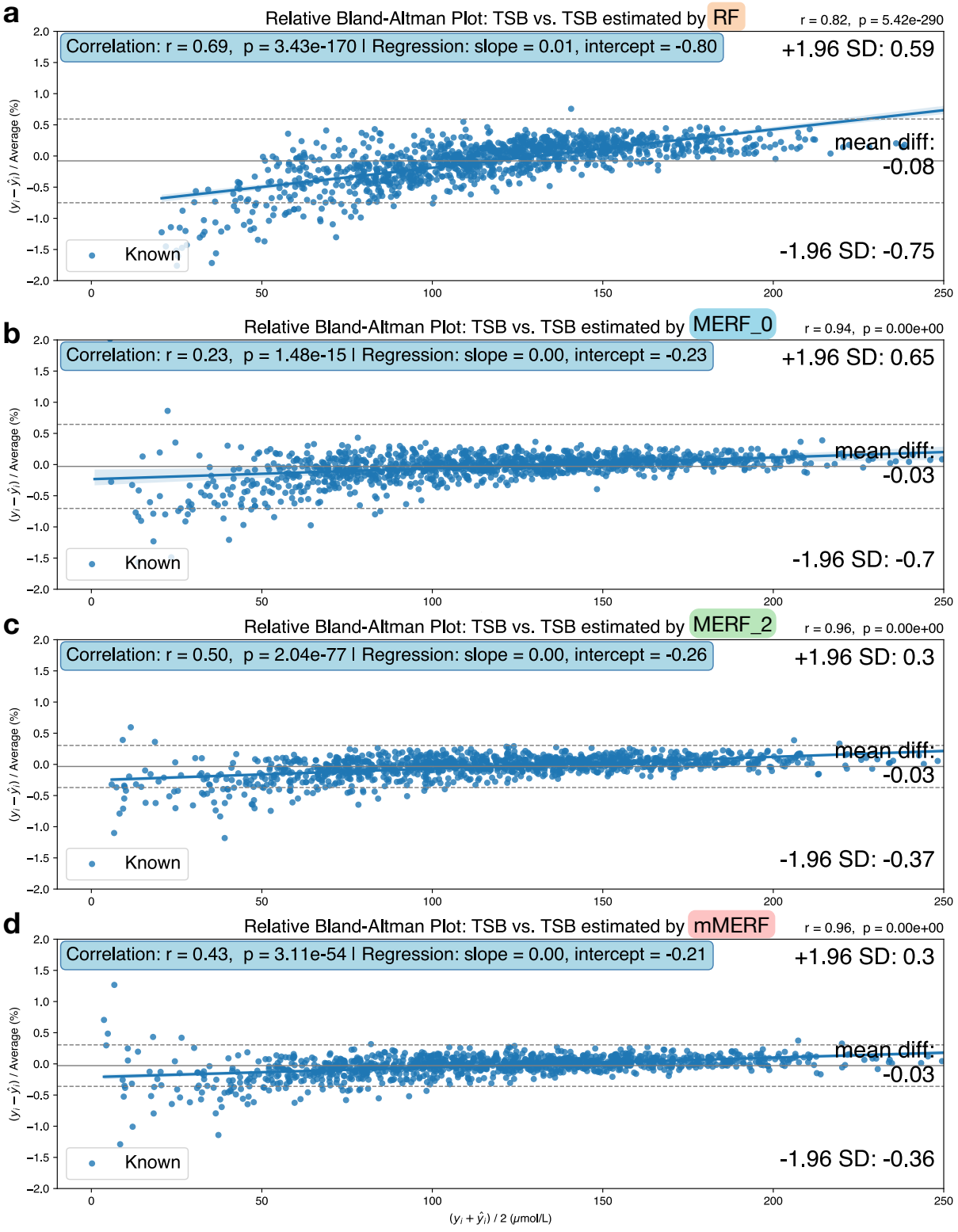

Figure A.9: Relative Bland-Altman plots of the mean against the differences between observed (y_i) and estimated ($\hat{y}_i$) TSB levels in percentage values in the training set. (**a**) RF_{base} model. (**b**) MERF_0 model. (**c**) MERF_2 model. (**d**) mMERF model. The r and p on the upper right of each subplot indicate the Pearson correlations and associated p-values between the real TSB levels and the model-estimated TSB values.

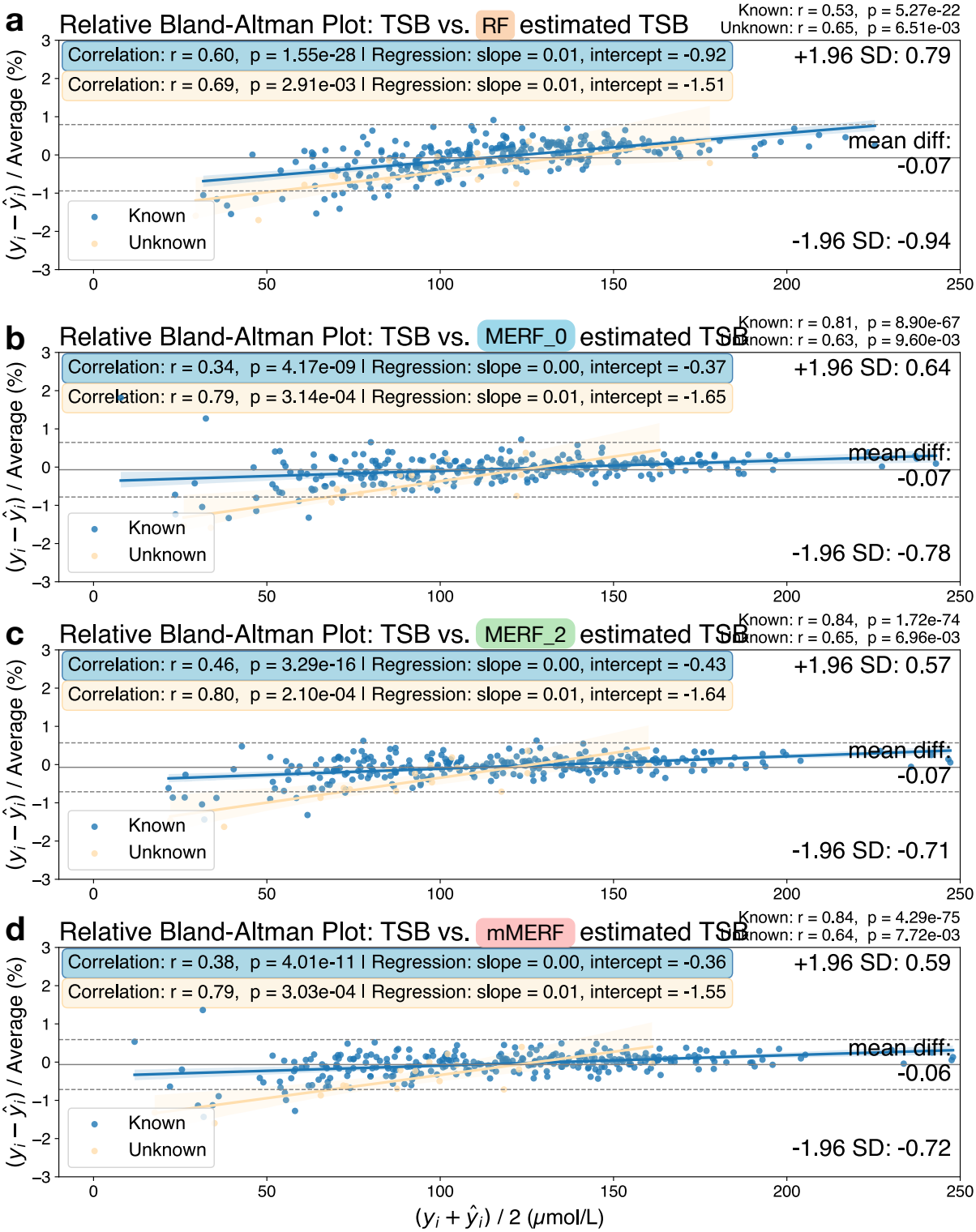

Figure A.10: Relative Bland-Altman plots of the mean against the differences between observed (y_i) and estimated ($\hat{y}_i$) TSB levels in percentage values for both samples of "known" patients (blue) and samples of "unknown" patients (beige) in the test set. (**a**) RF_{base} model. (**b**) $MERF_0$ model. (**c**) $MERF_2$ model. (**d**) mMERF model. The r and p on the upper right of each subplot indicate the Pearson correlations and associated p-values between the real TSB levels and the model-estimated TSB values.

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Traitement de données massives et apprentissage automatique explicable dans les unités de soins intensifs néonatals

Mot clés : Prématurité, unités de soins intensifs néonatals, hyperbilirubinémie, sepsis, variabilité cardiaque, traitement des signaux physiologiques, apprentissage automatique, système d'aide à la décision clinique

Résumé : Les nouveaux-nés prématurés sont vulnérables à des complications comme l'hyperbilirubinémie néonatale et le sepsis tardif (LOS), posant des défis importants dans les unités de soins intensifs néonatals (USIN). Malgré les avancées en matière de soins, la détection précoce et la gestion efficace de ces affections restent complexes. Cette thèse, basée sur l'étude CARESS-Premi (NCT01611740), vise à développer des techniques avancées de traitement des données et des modèles interprétables d'apprentissage automatique afin d'améliorer la prise de décision en USIN, via des systèmes de surveillance non invasifs, continus et en temps réel.

Les principales contributions comprennent : (i) une chaîne optimisée de traitement des signaux pour l'analyse ECG en conditions réelles, adaptée aux USIN ; (ii) un modèle mathématique patient-spécifique pour la caractérisation de la dynamique postnatale de la bilirubine, avec des paramètres comme biomarqueurs potentiels pour détecter les comorbidités associées ; (iii) une estimation non invasive de la bilirubine utilisant des modèles d'apprentissage automatique à effets mixtes intégrant l'analyse de la variabilité de la fréquence cardiaque (HRV) et des informations physiologiques ; (iv) des modèles pour la détection précoce du LOS via l'analyse de la HRV ; (v) la conception, le déploiement et l'évaluation préliminaire d'un système d'aide à la décision clinique (CDSS) *on-the-edge*, intégrant du traitement des signaux en quasi-temps réel et des modèles d'inférence dans un contexte USIN. Ces résultats démontrent le potentiel du traitement avancé des signaux physiologiques combiné à l'apprentissage automatique pour optimiser les soins néonatals.

Title: Massive data processing and explainable machine learning in neonatal intensive care units

Keywords: Preterm, neonatal intensive care units, hyperbilirubinemia, sepsis, heart rate variability, physiological signal processing, machine learning, clinical decision support system

Abstract: Preterm infants are highly vulnerable to complications such as neonatal hyperbilirubinemia and late-onset sepsis (LOS), which pose significant challenges in Neonatal Intensive Care Units (NICU). Despite advancements in neonatal care, early detection and effective management of these conditions remain difficult. Based on the CARESS-Premi project (NCT01611740), the dissertation aims to develop advanced data processing techniques and interpretable machine learning (ML) models to enhance NICU decision-making and neonatal outcomes, by leveraging non-invasive, continuous and real-time monitoring systems.

The main contributions include: (i) an optimized automatic signal processing pipeline for real-life ECG analysis tailored to NICU; (ii) a patient-specific mathematical model for postnatal bilirubin dynamics characterization in preterm infants, with model parameters serving as potential biomarkers for detecting associated comorbidities; (iii) the knowledge-based non-invasive bilirubin estimation using mixed-effects ML integrating heart rate variability (HRV) analysis and physiological insights; (iv) ML models for LOS early detection using HRV analysis, proving timely alerts before clinical suspicion; (v) the design, deployment and preliminary evaluation of an on-the-edge clinical decision support system (CDSS) integrating quasi-real-time signal processing and ML models in a NICU setting. These results demonstrate the potential of combining advanced physiological signal processing with ML to optimize neonatal care.