FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A Data-Driven Framework for the Assessment and Mitigation of Clinical Measurement Disparities

Inês Alves Martins

Master in Bioengineering - Biomedical Engineering

Supervisor: Jaime dos Santos Cardoso, PhD

Co-supervisor: João Carlos Ramos Gonçalves Matos, MSc

Co-supervisor: Tiago Filipe Sousa Gonçalves, MSc

July 17, 2024

A Data-Driven Framework for the Assessment and Mitigation of Clinical Measurement Disparities

Inês Alves Martins

Master in Bioengineering - Biomedical Engineering

July 17, 2024

Resumo

Estamos perante uma era de desenvolvimento e proliferação de tecnologia a um ritmo extraordinário. A inteligência artificial é uma das áreas que capta mais atenção e já está integrada nas nossas vidas, incluindo na saúde. Existem modelos de *machine learning* que desafiam o desempenho dos clínicos em tarefas como o diagnóstico e o desenvolvimento de tratamentos inovadores. A área da saúde é um sistema complexo, mas alguns algoritmos também o são, com capacidades surpreendentes de reconhecimento de padrões. No entanto, os modelos podem refletir a subjetividade de quem os criou ou o viés introduzido nos dados.

Os dados médicos são registados pelos clínicos ou adquiridos por dispositivos médicos. Há evidências na literatura de que dispositivos médicos amplamente utilizados, como oxímetros de pulso, termómetros e eletrocardiógrafos, podem fazer diferentes medições entre subgrupos de pacientes. Foi mostrado que os oxímetros não medem igualmente a saturação de oxigénio no sangue em subpopulações com diferentes pigmentações de pele. Termómetros baseadas em tecnologia de infravermelhos podem ser outro caso de má calibração. Medições temporais (baseadas em infravermelhos) foram associadas a *odds ratio* significativamente menores de identificar febre em pacientes negros em comparação com medições orais, mas tal não se verificou em pacientes brancos. Especialmente nas Unidades de Cuidados Intensivos, onde os pacientes são mais instáveis, os dados médicos são recolhidos continuamente e podem influenciar as decisões dos clínicos.

Esta dissertação teve como objetivo desenvolver uma *framework* para a avaliação e mitigação do viés de dispositivos médicos em tarefas de *machine learning*. A oximetria e os termómetros de infravermelhos foram os casos de uso escolhidos. Uma versão ampliada do BOLD *dataset*, que contém pares de medições de oximetria de pulso (SpO_2) e de saturação do oxigénio no sangue arterial (SaO_2), foi usada no primeiro caso de uso. Um dataset com pares de medições de temperatura foi construído para o segundo. Uma primeira exploração deste *dataset* foi realizada calculando a prevalência de febre, *odds ratio* de febre e *odds ratio* de febre não identificada em pacientes com suspeita de infeção, comparando medições orais e temporais, em pacientes negros e brancos. Cada *dataset* foi dividido em dois subconjuntos: um sem a medida de referência e o outro sem a medida enviesada. Regressão Logística, Random Forest e XGBoost foram usados em três tarefas de previsão: mortalidade no hospital, pontuação SOFA respiratória nas 24h seguintes e aumento da pontuação SOFA geral. Os resultados de cada subconjunto foram comparados para identificar diferenças significativas entre eles. Por fim, três abordagens foram propostas para mitigar o viés.

A exploração do *dataset* de termometria mostrou resultados em direções opostas para pacientes negros e brancos, mas que não foram estatisticamente significativos. A *framework* desenvolvida permitiu uma melhor avaliação do viés nas medições. As abordagens para a mitigação do viés não foram globalmente vantajosas, embora algumas disparidades tenham sido reduzidas.

Concluindo, a *framework* desenvolvida foi projetada como uma ferramenta útil e facilmente aplicável a outros casos de uso. Esperamos trazer consciencialização sobre as desigualdades na saúde que afetam populações já de si vulneráveis e promover o desenvolvimento de soluções mais equitativas.

Abstract

We are witnessing times of technology development and proliferation at an extraordinary pace. Artificial intelligence is one of the fields that is capturing everyone's attention. It is already integrated into our lives, including healthcare. There are machine learning models that challenge clinicians' performance in tasks such as diagnosis and development of innovative treatments. Healthcare is a complex system but so are some algorithms, which have amazing capacities for pattern recognition. However, models can reflect the subjective thinking of their creators or the bias introduced in data.

Medical data is registered by clinicians or acquired by medical devices. There is evidence in the literature that widely used medical devices, such as pulse oximeters, thermometers, and electrocardiography machines, can return distinct measures across patient subgroups. Studies showed that oximeters measure blood oxygen saturation differently in subpopulations with different skin pigmentation. Infrared-based thermometers might be another case of miscalibration, as temporal (infrared-based) measurements were already linked to significantly lower odds of identifying fever in Black patients compared to oral measurements, but not in White patients. Especially in the Intensive Care Unit, where patients are more unstable, medical data is continuously being collected and might impact clinicians' decisions.

This dissertation aimed to develop a framework for the assessment and mitigation of medical device bias on downstream machine learning tasks. Oximetry and infrared-based thermometry were the chosen use cases. An extended version of the BOLD dataset, which contains paired pulse oximetry readings (SpO_2) and arterial blood gas measurements (SaO_2), was used in the first use case. A dataset with paired temperature measurements was built for the second. A first exploration of this derived dataset was performed by computing the fever prevalence, odds ratio of fever and odds ratio of hidden fever in patients with suspected infection, comparing oral and temporal measurements, in Black and White patients. Each dataset was divided into two subsets: one without the ground truth and the other without the biased measurement. Logistic Regression, Random Forest and XGBoost were used on three prediction tasks: in-hospital mortality, respiratory SOFA score in the next 24h and overall SOFA score increase. Results from each subset were compared to identify bias. Lastly, three approaches were proposed to mitigate the bias.

The exploration of the thermometry dataset showed results in opposite directions for Black and White patients but were not statistically significant. The developed framework allowed us to assess bias in the measurements in certain subgroups, especially in the blood-gas and oximetry study. The approaches for bias mitigation were not globally advantageous, although some disparities were reduced.

To conclude, the developed framework was designed to be a useful tool and easily applicable to other use cases. We hope to bring awareness to inequities in the healthcare systems that impact already vulnerable populations and to promote the development of more equitable solutions.

Agradecimentos
Acknowledgements

Gostaria de começar por agradecer ao meu orientador Professor Jaime Cardoso e co-orientadores Tiago Gonçalves e João Matos, pelo constante acompanhamento e por me motivarem a levar os desafios em frente. Em particular, ao Tiago e ao João, por estarem sempre à distância de uma mensagem, mesmo com os fusos horários complicados de gerir. I would also like to thank Leo A. Celi for his advice and for opening new horizons and opportunities. To the Laboratory for Computational Physiology team, thank you for the warm reception during my short stay at MIT.

O meu percurso académico foi também repleto de associativismo. Neste ponto, não podia deixar de mencionar o Núcleo de Estudantes de Bioengenharia, que tanto me fez crescer e que me fez ver que o nosso esforço conjunto pode mover montanhas. Muito obrigada a todos os que me acompanharam e que também deram um pouco de si.

Para além disso, levo a Tuna Feminina de Engenharia da Universidade do Porto sempre no coração. Madrinha, maninhas, rebentos e queridas amigas, muito obrigada por acreditarem em mim e tornarem esta aventura mais feliz. Vivi Engenharia com euforia, a tocar guitarras. Foram ensaios infinitos, chamadas por Zoom e muita arrumação de trotinetes. Acima de tudo, foram momentos cheios de amizade na Faculdade de Engenharia.

Deixo uma palavra de apreço muito sentida a todos os que viveram comigo as aulas, as festas, as praxes. Foi também convosco que passei dos melhores momentos desta vida académica, entre risos e lágrimas sentidos. A ti, querida Casquinhos, que, desde o primeiro dia, me deste (literalmente) a mão, agradeço do fundo do coração por tantos anos memoráveis ao teu lado quanto o número de vírgulas nesta frase.

Por fim, gostaria de agradecer à minha família por acreditarem sempre em mim, me ajudarem a crescer e me fazerem ver como este percurso foi bonito. Deixo um enorme obrigada a vocês, Mãe, Pai e Irmão, por me acompanharem e apoiarem nas minhas decisões e aventuras, mesmo quando ainda só sabia explicar que Bioengenharia era um curso abrangente. Olho para trás com orgulho de todas as vezes em que agarrei os desafios e as oportunidades que surgiram, mesmo que isso implicasse fazer com que o dia tivesse mais de 24 horas.

A ti, Avó Laura, a quem cedo da escola privaram, dedico este trabalho.

Inês Martins

"Se tanto me dói que as coisas passem
É porque cada instante em mim foi vivo
Na luta por um bem definitivo
Em que as coisas de amor se eternizassem."

Sophia de Mello Breyner Andresen

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Inteligence
AUROC	Area Under the Receiver Operating Characteristic Curve
CART	Classification And Regression Trees
CI	Confidence Interval
CITI	Collaborative Institutional Training Initiative
CRD	Collaborative Research Database
DL	Deep Learning
DT	Decision Tree
EHR	Electronic Health Record
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
GBM	Gradient Boosting Model
GCS	Glasgow Coma Scale
ICU	Intensive Care Unit
IQR	Interquartile Range
IR	Infrared
LASSO	Least Absolute Shrinkage and Selection Operator
LoS	Length of Stay
LR	Logistic Regression
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
MLP	Multilayer Perceptron
OR	Odds Ratio
R^2	Coefficient of determination
RF	Random Forest
RMSE	Root Mean Square Error
SAD	Sepsis-Associated Delirium
SD	Standard Deviation
SDG	Sustainable Development Goals
SOFA	Sequential Organ Failure Assessment
SVM	Support Vector Machine
WHO	World Health Organization
XGBoost	eXtreme Gradient Boosting

Chapter 1

Introduction

1.1 Context

1.1.1 Artificial Intelligence and Healthcare

Technology has developed and spread rapidly in the last decades with one field particularly in vogue: artificial intelligence (AI). AI is a field of computer science that tries to recreate human thinking with computation, enabling machines to learn and solve problems or even synthesize human language. Machine learning (ML) is a subfield characterized by models that can detect patterns in structured data and make classifications or predictions. Deep learning (DL) is a subfield of ML that takes one step further in complexity, allowing machines to learn more abstract patterns using artificial neural networks. When these neural networks become capable of generating new data, they are classified as Generative AI [16, 17].

AI is increasingly present in many fields of our lives and has already begun to be applied to healthcare, such as in diagnosis and treatment guidance tasks [18]. Despite the several challenges of AI usage in clinical practice, the advantages of successful solutions are undeniable, capturing everyone's attention. There are already algorithms that challenge radiologists in tumour identification; guidance of cohorts construction for costly clinical trials [18]; and prediction of patient-reported pain based on knee X-rays more equitably than physicians [19]. Big tech companies are also investing in this field, like Google working on prediction models for life-threatening conditions, such as sepsis and heart failure [20].

1.1.2 Through a Biased Lens

Even though we see technology as a mathematical and objective tool, it can reflect the subjective thinking of its creators. That is why we see biased results in certain technologies even today. For example, photographic films were calibrated for white skin so, until 1970, when some adjustments were finally made, images of darker-skinned people looked distorted [21]. Some medical devices, such as pulse oximeters, also rely on light sensors and seem to have similar calibration flaws [22]. They estimate arterial oxygen saturation by measuring light absorption at two light wavelengths

(660nm - red and 940nm - infrared) of oxyhemoglobin and deoxyhemoglobin in capillary blood, but this physical principle can be independently affected by skin tone [23]. There is evidence in the literature that these devices measure the blood oxygen saturation differently across subpopulations [21, 24]. Infrared-based thermometers might be another case of miscalibration. There is little research on the topic but Bhavani et al. [25] linked temporal (infrared-based) measurements to significantly lower odds of identifying fever in Black patients at multiple fever cutoffs. However, this association was not observed in White patients.

Medical devices, such as oximeters, continuously collect relevant data for clinicians to take action in the medical environment, especially in the Intensive Care Unit (ICU), where patients are more unstable. By collecting and storing clinical data, ML models can learn that information and provide effective and objective diagnoses or guidance. However, if measurements are already biased, these technologies might learn them and lead to disparities in care. Equity must be ensured to prevent discrimination and improve health outcomes across diverse populations.

1.1.3 Fairness and Equity

The concepts of equality and equity, specifically in the healthcare environment, are essential to this study and can be defined as follows:

- **Equality**: means giving the same treatment to different people or groups of people, from a one-size-fits-all perspective. Although this concept is not the same as equity, its evaluation is crucial to achieving greater equity [26];
- **Equity**: the World Health Organization (WHO) defines health equity as "the absence of unfair, avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically, or geographically or by other dimensions of inequality (e.g. sex, gender, ethnicity, disability, or sexual orientation)" [27].

It must be noted that they are distinct concepts. To eliminate health disparities and guarantee equity, the different circumstances among people have to be taken into account so resources can be allocated properly [28]. For example, let us consider an AI model to predict potential dermatological conditions. An egalitarian model would perform similarly in all subpopulations, while an equitable one would prioritize performance improvements in groups with pre-existing health disparities so that each group can attain its full potential [1].

Although these concepts are extremely important to guarantee an effective action and correct assessment of the progress toward health equity, different entities still define them differently. This is one of the current barriers to improvement [29].

1.2 Motivation

Technology and AI have already shown significant potential for improvement in healthcare. While there is ample room for growth and development, other priorities cannot be forgotten. Medical

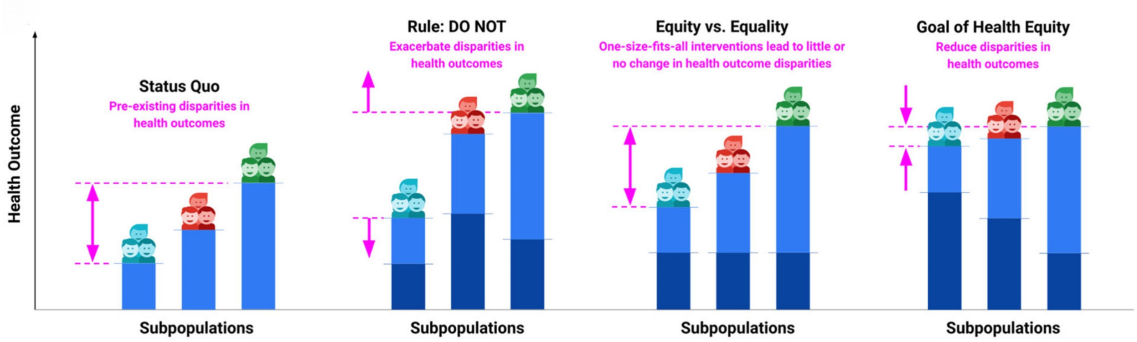


Figure 1.1: A visual representation of different approaches and their consequences on health outcomes. The light and dark blue bars represent the pre-existing health outcomes and the contribution of each approach to those outcomes, respectively (from [1]).

devices provide clinicians with essential information to evaluate the patient's status, but they are often designed and validated in populations lacking diversity, which can lead to measurement disparities. This bias might impact the clinician's immediate decision and downstream machine learning models based on the acquired data and corresponding conclusions. This issue must be highlighted to address proper solutions. However, to showcase measurement bias and evaluate health outcomes, researchers and developers require access to clinical data.

Closing our eyes to disparities will only amplify current discrimination. A device or algorithm that performs differently across subpopulations and exacerbates social disparities is not an improvement. We need solutions that promote more equitable healthcare.

1.3 Institutional Partnerships

This dissertation was a partnership between the Faculty of Engineering of the University of Porto and the Laboratory for Computational Physiology (LCP) at the Massachusetts Institute of Technology (MIT). The straight connections with developers of the PhysioNet databases and experts on the topic were truly advantageous.

A clinically diverse and certified board gave advice and validated several steps of this dissertation, such as the building process of the thermometry-linked dataset (Section 4.1.2) and the medical feature selection, which would be used in the designed framework for the assessment and mitigation of medical device bias on downstream machine learning tasks (Section 4.3.1.2).

The partnership opened doors to a short-period stay at MIT, which allowed me to better understand the motivations of this work. This travel enabled participation in the MIT-MGB AI Cures Conference 2024, in Boston, and in the Duke Critical Care Datathon 2024, at Duke University.

1.4 Objectives

The objectives of this dissertation are threefold:

- First, curate a dataset with paired thermometry measurements and additional medical information;
- Second, create a framework for the assessment of medical device bias on downstream machine learning tasks, using oximeters and infrared-based thermometers as two use cases;
- Third, propose solutions to mitigate the problem and re-evaluate the machine learning outcomes.

Furthermore, this dissertation aims to bring awareness to inequities in the healthcare systems that are (made) invisible but have a huge impact on already vulnerable populations.

1.5 Main contributions

The main contributions of this dissertation are:

- The curation of a temperature-paired dataset, facilitating research on similar topics. This dataset will be extended with other databases to improve robustness and population diversity, with the objective of making it freely available.
- The statistical assessment of temperature bias in septic patients from MIMIC-IV. The preliminary results were presented at three conferences:
  - Science Under '5, a pitch competition organized within the XV Symposium on Bioengineering, in Porto;
  - Poster session at the MIT-MGB AI Cures Conference 2024, in Boston;
  - Poster session at IJUP'24 – 17th Young Researchers Meeting of the University of Porto.
- A contribution to the paper submitted to JAMA, titled "Racial Differences in Temporal Thermometry Associated with Delayed Sepsis Bundle Care". It is related to the methodology described in Section 4.2.
- The development of a framework for the assessment and mitigation of clinical measurement disparities in three downstream ML tasks, using oximetry and thermometry as initial use cases. This framework was designed to be a useful tool and easily applicable to other use cases. It led to the submission of a paper for the Third Workshop on Applications of Medical AI (AMAI 2024), titled "Evaluating the Impact of Pulse Oximetry Bias in Machine Learning under Counterfactual Thinking".

1.6 Sustainable Development Goals

The Sustainable Development Goals (SDG) [2] are a set of 17 objectives adopted by the United Nations in 2015, each with specific targets to be achieved by 2030. The project tackles global

challenges, such as poverty, hunger, healthcare, and climate change, among many others. All the 17 Goals are presented in Figure 1.2. This dissertation is integrated into two of them: Goal 3 - Ensure healthy lives and promote well-being for all at all ages, and Goal 10 - Reduce inequality within and among countries[1].

Figure 1.2: The 17 Sustainable Development Goals (from [2]).

1.6.1 Goal 3: Ensure healthy lives and promote well-being for all at all ages

The third SDG was developed to promote global health coverage. Progress was made in several aspects, such as the decrease of under-5 mortality and the effectiveness of HIV treatment. However, the COVID-19 pandemic and other worrying circumstances caused setbacks, as in childhood vaccination. Continuous investment in healthcare is crucial to maintaining the conditions that allow countries to develop and apply their strategies, even in and after adverse situations. Access to adequate healthcare, medicines and vaccines for all is one of the specific targets but it can only be achieved if firstly inequities are highlighted. Hence, individuals and entities are aware of the problem and capable of acting accordingly.

1.6.2 Goal 10: Reduce inequality within and among countries

There were already clear differences within and among countries at social and economic levels and the pandemic was a major contributor to its increase. The opposite way should be followed with more equitable resource distribution, discrimination sources identification and support of vulnerable populations. It is brutal to think that one in six people worldwide was somehow discriminated.

[1]The content of this publication has not been approved by the United Nations and does not reflect the views of the United Nations or its officials or Member States.

A successful long-term development is based on diversity respect and inclusion in its many forms (e.g., sex, age, sexual orientation, ethnicity, religion) and every sector (e.g., social, economic and political).

1.7 Dissertation Structure

This dissertation is organized into six chapters. The present Chapter 1 contains a brief contextualisation on the topic, the motivation, the objectives, the relevance to the Sustainable Development Goals, and the main contributions of this work. Chapter 2 illustrates current applications and relevant research on machine learning in healthcare, and contains background knowledge about some of the steps of a machine learning pipeline and particular models. Chapter 3 better explains the two medical devices considered in this study: oximeters and thermometers. The methodology is described in Chapter 4 and the results and discussion are presented in Chapter 5. Chapter 6 is a conclusion on the results and a reflection on the initial goals and future work. This document ends with a set of appendices (Appendices A and B), containing more detailed information about the steps for the dataset curation and the complete results of the ML framework for both use cases.

Chapter 2

Machine Learning for Healthcare

This chapter outlines the methodologies applied to Electronic Health Record (EHR) data to support clinical decisions, providing background knowledge on the key steps of a ML pipeline and relevant models.

2.1 Is machine learning transforming patient care?

The power of AI is becoming undeniable. In healthcare, algorithms can be trained to predict patients' outcomes, such as the probability of developing a certain disease during their hospitalization or the duration of that stay, and guide clinical decisions. To become a valuable tool, they are expected to surpass the clinician's performance and bring new insights. For example, Pierson et al. [19] showed that a knee X-ray algorithm could predict patient-reported pain without pre-existing demographic and socioeconomic bias, introduced by physicians. These predictions could decrease inequalities in treatment access, like arthroplasty. In addition, the model's complexity has to be balanced and developed in parallel with explainability, to gain the trust of the clinicians and the remaining entities related to the healthcare system.

Logically, one of the outcomes of utmost concern is mortality. What would be the impact of an accurate prediction of in-hospital mortality from patients' health records? Several authors have been using EHR data to build models with this prediction goal. Bao et al. [13] predicted sepsis patients' in-hospital mortality based on age, sex, weight, vital signs, laboratory values, advanced cardiac life support, and accompanying diseases. 21,680 patients were included in the study. Seven models were used, firstly without parameter optimization: Support vector machine (SVM), Decision Tree (DT) Classifier, Random Forest (RF), Gradient Boosting (GBM), Multilayer Perceptron (MLP), eXtreme Gradient Boosting (XGBoost) and Light Gradients Boosting (LightGBM). LightGBM had the best performing metrics in the train and test sets: mean F1 scores of 0.909 and 0.906, respectively, and Area Under the Receiver Operating Characteristic Curve (AUROC) up to 0.85. The remaining results are presented in Table 2.1. After parameter tuning, LightGBM achieved AUROC values of 0.99 and 0.96 in the train and test sets, respectively.

Table 2.1: Train and test F1-scores and AUROC obtained in [13] with SVM, DT, RF, GBM, MLP, XGBoost, and LightGBM.

Metric		SVM	DT	RF	GBM	MLP	XGBoost	LightGBM
Train	F1-scores	0.649	0.786	0.886	0.899	0.644	0.896	0.909
	AUROC	0.70	0.75	0.82	0.85	0.82	0.84	0.86
Test	F1-scores	0.649	0.803	0.891	0.901	0.625	0.892	0.906
	AUROC	0.75	0.75	0.81	0.85	0.82	0.84	0.85

Another study [3] focused on patients with cardiac arrest, which is an event with a high mortality rate. The demographics, comorbidity, vital signs, laboratory test results, scoring systems, and treatment information on the first day of ICU admission of 1,722 subjects were considered. Feature selection was performed using the least absolute shrinkage and selection operator (LASSO) regression model and the XGBoost, in the training set, followed by a multivariate logistic regression (LR) analysis, culminating in different prediction models. The features selected both by LASSO and XGBoost were the following: age, heart rate, respiratory rate, bicarbonate, SpO_2, temperature, simplified acute physiology score III (SAPS III) score and Glasgow coma scale (GCS) score. LASSO, XGBoost and multivariate LR models outperformed the National Early Warning Score 2 (NEWS 2) model, which is a scoring system for potential disease deterioration detection that provides preventive measures. The best AUROC in the validation set was obtained using the LASSO model, reaching a value of 0.80 (Figure 2.1), with the remaining models having similar but slightly lower results.

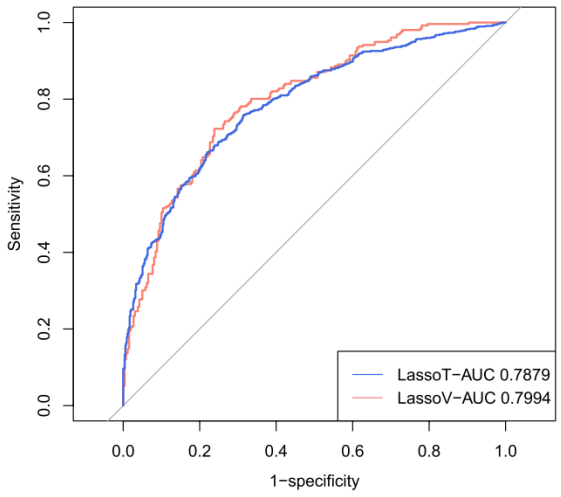


Figure 2.1: ROC Curves of the LASSO model in the training set (blue) and validation set (red) (from [3]).

An accurate diagnosis is a primary factor to avoid complications. However, when it does not occur in septic patients, they can develop sepsis-associated delirium (SAD). Zhang et al. [4] built a ML model for early prediction of this event. 14,620 and 1,723 patients were selected from MIMIC-IV [30] and eICU-CRD databases [31], respectively. LASSO regression was used to select features, which were included in the following groups: demographics; type of initial

ICU admission; vital signs; laboratory test results; Sequential Organ Failure Assessment (SOFA) and GCS scores; comorbidities; mechanical ventilation, continuous renal replacement therapy, vasopressors, sedatives; ICU length of stay; 28-days ICU mortality; diagnosis time for delirium and sepsis. The considered prediction models were LR, SVM, DT, RF, XGBoost, K-Nearest Neighbours, and Naïve Bayes. XGBoost had the best results and showed higher robustness and stability across different prevalence rates. The internal and external validation AUROCs were approximately 0.79 and 0.70, respectively (Figure 2.2). All remaining models had values above 0.69 and 0.61, for the internal and external validation AUROCs, respectively.

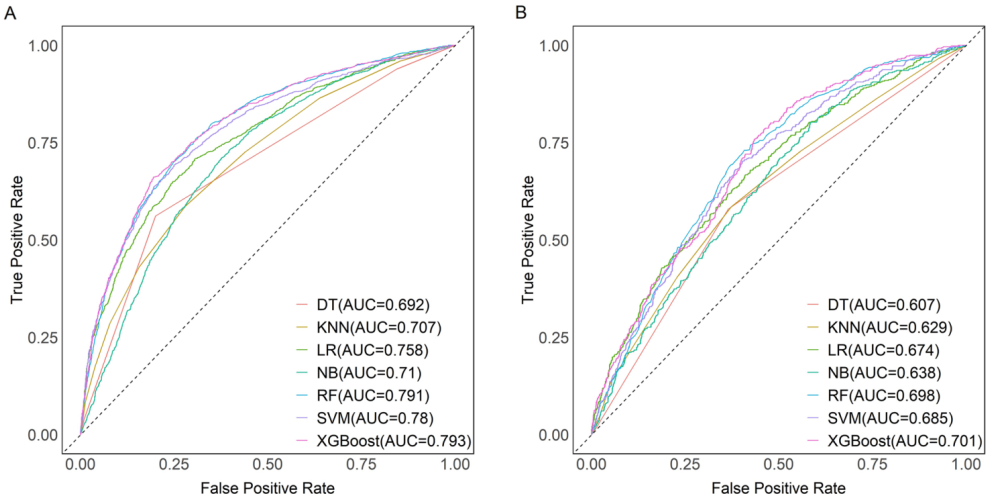


Figure 2.2: ROC Curves of the machine learning models on the internal (A) and external (B) validation sets (from [4]).

Appropriate healthcare is strongly influenced by hospital management and resource planning. Thus, the prediction of the patient's ICU length of stay (LoS) has also been a topic of research. Hempel et al. [15] included demographic, administrative, vital signs, and laboratory information from the first 24h of 41,473 ICU stays in their study. The chosen ML models were logistic regression, SVM, random forest and XGBoost. Both RF and XGBoost are widely used and show good performance in this prediction task [32, 33, 34]. In the first approach, LoS classification was binarized as "short" or "long". Although models' performances were not significantly different, RF achieved better results in almost every metric (accuracy = 0.81, F1-score = 0.44, AUROC = 0.80). Considering the regression prediction, SVM showed better results in terms of mean absolute error (MAE = 1.68) and mean absolute percentage error (MAPE = 48.37) and RF had the best root mean square error (RMSE = 2.81) and coefficient of determination (R^2 = 0.24). Kakadiaris [35] also studied LoS prediction, using the MIMIC-IV dataset and XGBoost as the prediction model. The task was binarized into short and extended stays. Demographics, vital signs, medication, and laboratory information were included. There were three approaches to dealing with missing data: using data with the missing values, inputting the feature's mean or inputting its median. The accuracy results were between 83% and 82% and AUROC between 0.85 and 0.86. Race and insurance were considered sensitive attributes, but no statistical analysis was published to compare

the different subgroups' results properly.

Going further in complexity, Beaulieu-Jones et al. [5] trained deep learning models with the purpose of understanding if they could improve patient risk stratification. There was the hypothesis that these models are merely learning the decisions already made by the clinicians, in the training part, which may not translate into better clinical decisions to new problems and health status. The prediction tasks were in-hospital mortality, 30-day readmission rate and prolonged LoS. Three subsets of patient features were considered: 1) demographic data only, 2) demographic data and information at admission, and 3) demographic data, information at admission and charges during the first day of admission. Models using subsets 1 and 2 were built with a LR due to the reduced amount of features and the last one utilised a stacked recurrent neural network. The third feature subset was the one with closer but still worse AUROC results than studies that included all EMR available data: 0.89 to in-hospital mortality, 0.71 to 30-day readmission rate and 0.82 to prolonged LoS (Figure 2.3). However, it may lead us to conclude that this deep neural network is not powerful enough to make predictions based only on the given information. Is it possible to improve performance at a point where we trust deep learning models to guide clinicians in individual decisions?

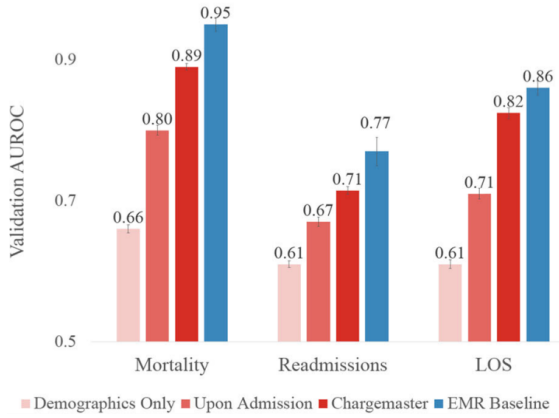


Figure 2.3: Comparison of mortality, readmission and LoS performance AUROC on the validation data across feature subsets (from [5]).

Rajkomar et al. [6] showed promising results. The complete patients' raw EHR data (including free-text notes) from two US academic medical centers (216,221 hospitalizations from 114,003 patients) were represented in the Fast Healthcare Interoperability Resources (FHIR) format. For predicting in-hospital mortality, 30-day unplanned readmission, prolonged LoS, and all of a patient's final discharge diagnoses based on this feature format, the AUROCs across sites were 0.93–0.94, 0.75–0.76, 0.85–0.86 and 0.90 (frequency-weighted AUROC), respectively. These results surpassed traditional models used in clinical settings, in a wide range of tasks, showing better deployment probabilities. Figure 2.4 shows the AUROC results of deep learning and baseline models regarding the in-hospital mortality task.

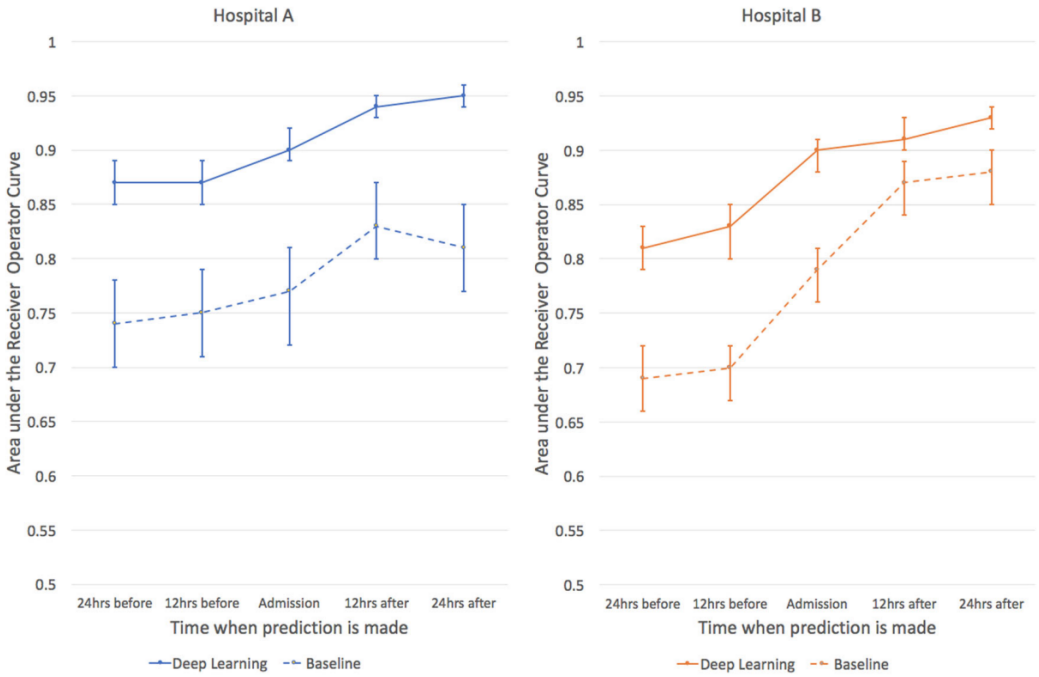


Figure 2.4: Comparison of mortality performance AUROC between deep learning and baseline models at different times before and after hospital admission. Hospital A (left) corresponds to the University of California, San Francisco (UCSF) and Hospital B to the University of Chicago Medicine (UCM). The error bars represent the 95% confidence interval (CI) (from [6]).

2.2 Unbiased Machine Learning: are we there yet?

The impact of ML solutions in healthcare is still limited, due to the complexity of the health environment and how algorithms are built and validated. All pipeline steps must be carefully examined to prevent the introduction and propagation of bias [36, 37]. Starting from the basis, the developer must guarantee that data is diverse and representative of the real world. When this requirement is missing, weights should be applied to samples of the minority groups to overcome the problem. Feature selection is another step that requires caution. A study made by Obermeyer et al. [38] concluded that an algorithm widely used in United States hospitals was less likely to refer Black people than White people to more personalized treatment programs, in situations of similar severity of illness. These inconsistencies were produced because the model was based on health care cost, instead of actual illness, and the preceding unequal access to treatments was propagating bias in forward predictions. More relevant variables should be selected. However, in Obermeyer's words: "the hard part is: what is that other variable? How do you work around the bias and injustice that is inherent in that society?" [39]. Continuous algorithm validation is needed [37]. By reorienting ML pipelines, achieving greater results and going forward into deployment will be possible. The advantages of an effective AI implementation in the healthcare environment are clear: diminishment of possible biased decisions, improvement of healthcare professionals' work conditions by offloading cognitive tasks, health system efficiency and overall population health results [40]. We must develop ML solutions for and not on healthcare.

2.3 Machine Learning models

Machine Learning has four subcategories: supervised, unsupervised, semi-supervised and reinforcement learning. A model is supervised when labeled training data is used to make outcome predictions of classification or regression problems. Unsupervised learning uses unlabeled data to find patterns that make clustering or aggregation possible. As the name suggests, semisupervised learning is a middle ground between the previous two: the input data is partially labeled. In reinforcement learning, an agent tries to find the optimal way to achieve a certain goal, maximizing the cumulative rewards through trial and error.

In this thesis, we focus on supervised models. Therefore, some relevant models are explained in more detail.

2.3.1 Logistic regression

LR is a supervised ML algorithm and is used for binary classification. Nevertheless, adjustments can be made in order to perform multiclass classification. It is widely applied in situations where data is prone to biased predictions when different groups are being analysed [41]. The LR algorithm first computes the natural logarithm of the odds (ratio between the probability of a sample belonging to one class and the probability of belonging to the other), commonly known as the log odds (Equation 2.1). It is written as a regression function of the predictors X_i, with w_i being the coefficients and w_0 the bias term/intercept.

$$\ln \frac{P(Y=0 \mid X)}{P(Y=1 \mid X)} = w_0 + \sum_i w_i X_i \tag{2.1}$$

Considering Equation 2.1 and some mathematical rearrangement, the final LR equation can be written as:

$$P(Y=1 \mid X) = \frac{1}{1+\exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)} \tag{2.2}$$

The general idea is to calculate and map these probabilities with the sigmoid (or logistic) function (Figure 2.5).

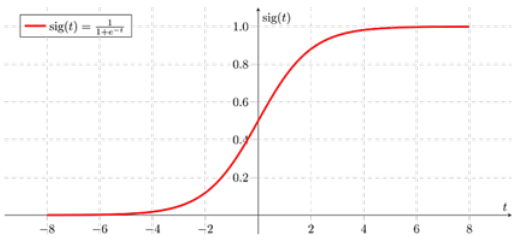


Figure 2.5: Sigmoid function (from [7]).

The three types of LR models are:

- Binary logistic regression: when the possible outcomes are dichotomous;
- Multinomial logistic regression: when there are more than two possible unordered outcomes;
- Ordinal logistic regression: when there are more than two possible outcomes but with a defined order.

To ensure a correct application of this model, we must be aware of its assumptions: the observations are independent, the output is dichotomous, variables and log odds have a linear relationship and the dataset must be free of outliers. When applying an LR model to a multinomial situation, the SoftMax instead of the sigmoid function is used to predict each probability. Although a high number of features may result in overfitting, regularisation can help prevent this issue.

Some examples of LR successful applications are the evaluation of the female gender as an independent predictor of increased operative mortality after coronary artery bypass graft [42]; or the assessment of the association between cholesteryl ester transfer protein *TaqIB* variant and coronary artery disease [43].

2.3.2 Binary Decision Tree

A DT is a non-parametric method which can be used for regression and classification problems. The algorithm consists of successive partitionings each containing a simple binary prediction model. These individual models are called "decision nodes", the two possible results are called "branches" and the outcomes are the "leaves", in analogy to the branching structure of a tree (Figure 2.6) [44]. This algorithm's main advantages are its simplicity, explainability and easy visualisation. Although it is very sensitive to changes in the dataset, ensemble methods can mitigate that problem, as explained next.

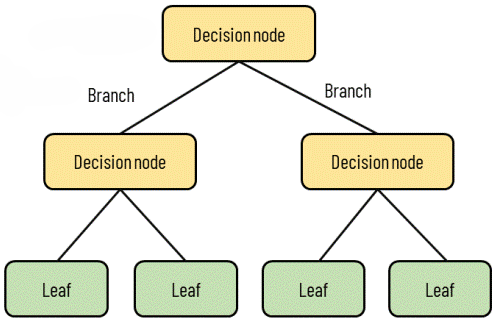


Figure 2.6: Decision tree structure (from [8]).

2.3.3 Random Forest

Before describing the following models, a few more concepts must be understood. An ensemble model consists of training several ML models independently and then giving an overall prediction by averaging, voting or weighting the previous individual predictions. Bagging - bootstrap

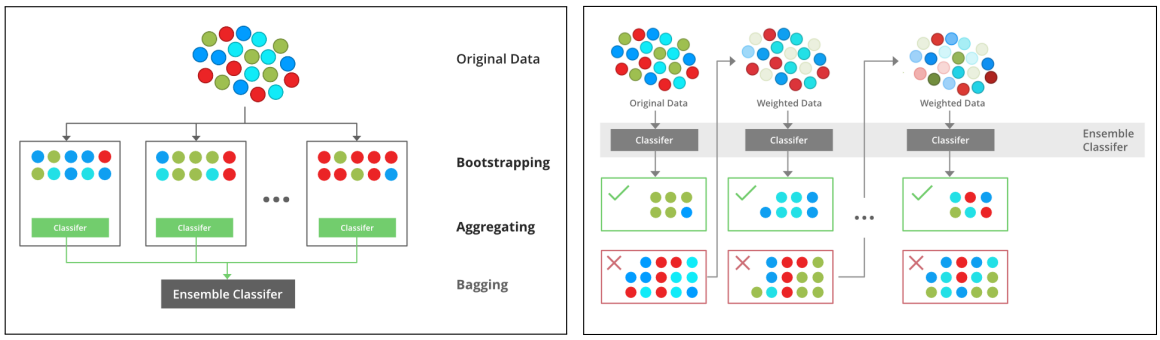


Figure 2.7: Ensemble models: bagging (left) and boosting (right) (from [9]).

followed by aggregation - is an example of an ensemble method. Data is randomly divided into subsets (bootstrap), each is used to train parallel models and, finally, overall output is calculated by averaging or voting (aggregation). Boosting is another ensemble model. In this case, models are trained sequentially and the information from the previous model is used to guide the learning in the next one. The contribution of the individual results to the overall output is weighed based on each performance. The structures of these two ensemble models are shown in Figure 2.7. Bagging helps reduce overfitting (variance) while boosting reduces bias.

Random Forest is based on the bagging method. The rows and features of the dataset are randomly sampled in subsets to train multiple decision trees, and the final result is calculated by voting or averaging all the outputs, in the case of a classification or a regression task, respectively [45].

2.3.4 XGBoost

Extreme Gradient Boosting, commonly known as XGBoost, is an optimised distributed gradient boosting library [46]. A Gradient Boosting model, which uses a boosting approach, inputs previous residual errors in the following predictors as labels to promote consecutive error-fitting [47]. There is a wide variety of loss functions to choose from, making the model adjustable to the researcher's needs. Specifically, the XGBoost algorithm is based on sequential Classification And Regression Trees (CART). A CART's leaf contains a score instead of only a decision value as a standard decision tree. Every independent variable is associated with a weight that is increased in the following predictions if the results are not correct. The XGBoost's complexity is controlled with both Lasso and Ridge Regression regularisations. Moreover, it has an embedded cross-validation method. XGBoost efficiency and flexibility make its methods capable of adapting to a wide range of supervised learning problems (regression, classification, ranking, and even user-customized) and achieving high performance. Additional advantages are the ability to handle missing values, getting closer to the real-world data; the suitability for large datasets, with a reasonable amount of processing time; and the possibility of returning feature importance, which improves interpretability. However, small datasets or using a high number of trees can result in overfitting. On the one hand, careful hyperparameter tuning will lead to great performances and

reduce overfitting. On the other hand, it is a time-consuming task and demands a deep knowledge of each parameter effect.

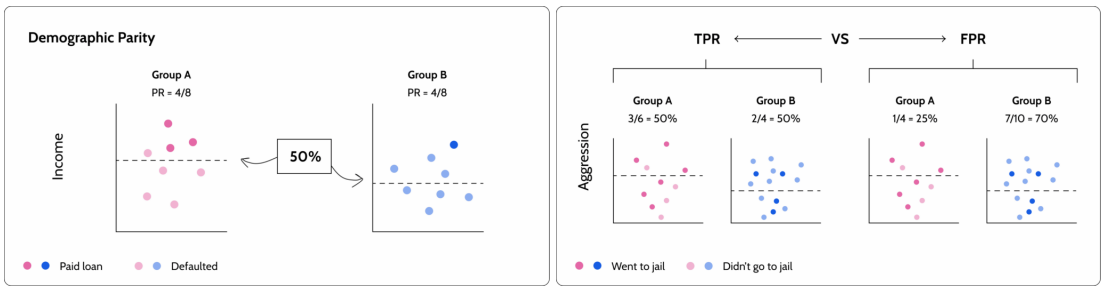
In Section 2.1, there are several examples where XGBoost outperforms other well-known machine learning models and achieves great performance.

2.4 Evaluation metrics

The model's performance can be evaluated with a wide range of metrics. Deciding the most relevant ones depends on the context and problem being tackled. However, they are complementary so conclusions should be based on several metrics [40, 48]. In a classification task, the common choices of fairness metrics are accuracy, precision, recall/sensitivity/true positive rate and specificity/false positive rate. F1 score is a derived metric from precision and recall, computed with $\frac{2\times precision\times recall}{precision+recall}$. The Area Under the Receiver Operating Characteristics (AUROC) is the integral of the sensitivity vs. specificity plot and represents the model's capability to separate two classes.

Additional metrics to be taken into account that reflect the equity in the results are:

- Demographic parity difference - difference between the largest and the smallest predicted positive rate, across all values of the sensitive feature(s). Lower values mean more similar selection rates. In Figure 2.8a, the predicted positive rate is 50% for both groups so the demographic parity difference is 0.
- Equalized odds ratio – corresponds to the smaller value between the true positive rate (TPR) ratio and the false positive rate (FPR) ratio. These ratios are calculated between the correspondent smallest and largest rates across all values of the sensitive feature(s). This metric being 1 means that the components of the confusion matrix are equal for all groups. In Figure 2.8b, the true positive rate ratio is 1 and the false positive rate ratio is 5/14 so the equalized odds ratio is 5/14.



(a) Demographic parity of groups A and B. (b) TPR and FPR of groups A and B.

Figure 2.8: Illustrative examples of demographic parity (a), TPR and FPR (b) calculations, applied in the Demographic parity difference and Equalized odds ratio computation (from [10]).

Regression problems can be calibrated using, for example [49]:

- Root mean square error (RMSE) - shows how far predictions are from true values using Euclidean distance:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{2.3}$$

- Coefficient of determination (R^2) - shows how well the regression approximates the predictions to the input:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.4}$$

where N is the number of samples, y_i is the true value, $\hat{y}_i$ is the predicted value and $\bar{y}$ is the mean of the true values.

2.5 What if?

Every action has an effect. But what would have happened if another decision was taken? This is counterfactual thinking.

Counterfactual models try to quantify the impact of specific changes in an environment, as an attempt to show a causal effect. In its simplest way, to build this type of model, it is only needed to change one feature's values and observe the possible changes in the outcome [50]. They can be applied to personalized medicine, by predicting the best individual outcomes in a range of possibilities. For example, more adequate and innovative treatments for immunotherapy can be developed by applying counterfactual models in the T-cell receptor sequence design [51].

Counterfactual AI is an interesting tool to link AI models' results with clinicians' knowledge and reasoning, increasing transparency and explainability. Additionally, they might improve patients, clinicians, regulators, and data scientists' trust in clinical AI classifiers and push model deployment.

There are two types of explanation methods: model-agnostic and model-specific, depending on whether they work with the input and output of the prediction or if the model structure is considered, respectively [50].

2.6 Summary

ML can be an amazing tool to support clinical decisions and should be incorporated into healthcare systems. The advantages are in front of our eyes: increased objectiveness in decisions, improvement of healthcare professionals' work conditions, advancements in system efficiency and overall population health results [40]. EHR data is one of the pillars for success. Several authors have been working on the prediction of certain outcomes, such as mortality, LoS and final discharge diagnoses. LR, RF, LightGBM and XGBoost models seem to be interesting approaches.

Nevertheless, all steps that lead to algorithm deployment must be carefully analyzed, from data acquisition to final validations, so bias is not introduced and continuously propagated [37, 38]. The

application of counterfactual thinking might increase transparency, explainability and trust. This is a continuous and difficult work because it often goes across several years and all entities involved in the process must be aligned with the same values [40].

Despite all challenges, we must develop ML solutions for and not on healthcare.

Chapter 3

Invisible Lines: Uncovering Issues in Medical Devices

This Chapter outlines the problems associated with miscalibrated medical devices, focusing on oximeters and non-contact thermometers and the corresponding consequences to patients' health.

3.1 Introduction

In Chapter 2, attention was brought to possible mistakes in machine learning pipelines. What if bias is introduced even before that? In healthcare applications, algorithms are trained with data at least partially collected by medical devices. Previous research has shown that widely used medical devices, such as pulse oximeters, thermometers, electrocardiography machines, and sphygmomanometers, can return different measures across patient subgroups [22]. Introducing biased values in a machine learning model will make it learn and propagate this issue. Concerns about oximetry and thermometry are presented in more detail and will be the focus of this dissertation.

Figure 3.1: Medical devices: oximeter (left) and temporal thermometer (right) (from [11] and [12], respectively).

3.2 Up and down: oxygen saturation measurements

Blood oxygen saturation is one of the most relevant parameters to assess health status, given that abnormally low levels compromise organ function. It can be measured by arterial blood gas (ABG), the "gold standard" but an invasive technique [52]. Pulse oximeters are also used to monitor blood oxygenation and were developed in 1972 as an alternative to invasive arterial oxygen saturation (SaO_2) measurements. It measures the peripheral oxygen saturation (SpO_2) by comparing oxygenated and deoxygenated hemoglobin's absorption of red and near-infrared light [22]. Its noninvasiveness, easiness of data collection, affordability, and nearly continuous and real-time acquisition rapidly made it a popular device. However, it is known that its original validation was not performed on a diverse population [21]. Accurate measurements are needed to guarantee a correct diagnosis and adequate treatment. An example that exacerbates its importance is silent hypoxia, which happened to be the only symptom in some COVID-19 patients. It is characterized by hypoxia (low oxygen levels in the tissues [53]) without dyspnoea (shortness of breath or difficulty breathing) [54].

Oximeters are class II medical devices and are under the Food and Drug Administration (FDA) 510(k) recommendations. Their performance evaluation studies should be conducted as described in ISO 80601-2-61:2017 *Medical Electrical Equipment — Part 2-61: Particular requirements for basic safety and essential performance of pulse oximeter equipment* [55] or equivalent method. One of the requirements is the validation in a real-world representative population, including a variety of age, sex and skin pigmentation (at least 15% of the subjects should have dark skin pigmentation).

K. Poorzargar et al. [56] published a systematic review on the measuring accuracy of pulse oximeters and noticed that only one of the considered studies took into account skin pigmentation. Most have a higher percentage of male subjects, with one reaching the extreme of an all-male population. Multiple studies were conducted to quantify this potential bias in pulse oximeters and the results are worrying. Sjoding et al. [24] found that Black patients experienced nearly triple hidden hypoxemia cases compared to White patients when using pulse oximetry measurements instead of arterial oxygen saturation in arterial blood gas. These discrepancies were associated with inequities in oxygen therapies, subsequently higher organ dysfunction scores and increased mortality rates among subpopulations [57]. Other authors showed overestimated SpO_2 values during hypoxia in dark-skinned subjects [58, 59]. One study tested six devices used at home and hospital settings by taking temperature measurements through neutral density and varying synthetic melanin filters. This strategy has the advantage of not relying on self-reported race and/or ethnicity. Performance varied significantly across different pulse oximeters. It was not possible to conclude that melanin does not affect SpO_2 values, so the authors' recommendations are to go further in calibration, theory and instrument design [60].

3.3 Hot and cold: temperature measurements

Monitoring and management of body temperature are usual procedures in medical environments because temperature values may help to understand patients' health status, such as the presence of an infection. In fact, more than two-thirds of patients admitted to the ICU have a fever at the moment of admission [61]. Not only hyperthermia (fever) but also hypothermia are associated with higher mortality rates in the ICU [62, 63]. The temperature thresholds for fever and hypothermia identification may vary according to the disease being studied. Regarding sepsis, they are established as 38.0°C and 36.0°C, respectively [64].

Sepsis is defined as a life-threatening organ dysfunction caused by a dysregulated host response to infection [65] and is the most common cause of death in critically ill patients in the United States [64]. If correctly identified, it can be treated with antibiotics, hemodynamic support and control of the source of infection, although this is a complex management [66]. Around $20.3 billion are spent annually on sepsis-related care, in the United States [64]. EHR might be key if measurement quality is ensured. Around 60% of these critical patients have fever at the time of admission [67] and 10-30% are hypothermic [68, 69, 70]. A meta-analysis of clinical trials shows that the mortality rates of patients with sepsis are around 22.2%, 31.2% and 47.3% to the ones with fever, normal temperature and hypothermia, respectively [71]. However, some studies show increased mortality rates associated with hypothermia but not with fever, in septic patients [72, 73]. Furthermore, it remains uncertain if the existing ICU severity scoring systems that take temperature into account consider adequate ranges for all subpopulations. Their calibration and predictive precision may differ significantly based on several patient demographics and diseases [74].

Core body temperature is defined as the organs' temperature, which can be measured by invasive methods. On the other hand, the temperature of the body surfaces, measured with non-invasive devices, is called "peripheral" [75]. Body temperature can be measured using contact (which can be invasive or not) or non-contact (infrared {IR}-based) thermometers. However, it must be noticed that not all devices reflect core body temperature with the same accuracy.

There are several types of temperature reading devices, which use different technologies and have variable accuracy.

Contact thermometers:

- Pulmonary artery temperatures are one of the gold standards for core temperature monitoring. However, this type of measurement is not recommended for routine temperature monitoring in the ICU because it is not practical, it is invasive and has the risk of infection [76].

- Rectal thermometers are very invasive and uncomfortable for the patient. It has been used as a core reference. However, a significant lag of rectal temperature behind other core sites was found in some studies, increased by fast temperature changes [76, 77, 78].

- Esophageal thermometers have a probe that measures the temperature in the distal third of the esophagus. The probe's location must be checked by a chest X-ray. Moreover, values may be inaccurate due to fluids or nasal breathing [76].

- Oral temperatures are measured by introducing a probe in the oral cavity. In general, they reflect core temperature very well. Nevertheless, a study showed that higher respiration rates can reduce oral temperatures by about 1°C per 25 resp/minute [77, 78].

Non-contact thermometers:

- Temporal artery thermometer uses an infrared sensor and measures the values in the superficial temporal artery. This region does not have mucous membranes and thermoregulatory stimuli almost do not influence the blood flow [77]. In addition, these devices allow a fast mass screening, reducing contact with potentially ill patients.

- Tympanic measurements also rely on infrared technology. However, the result highly depends on the probe's direction when inserted [77]. The waste caused by the use of disposable hygiene covers is an additional disadvantage [79].

Body temperature is a key factor in early identification and management of sepsis. Elevated temperature ($\geq$ 37.8°C) also turned out as one of the most prevalent indicators of COVID-19 – fever was detected in around 78% of the cases [79]. During that time, IR-based thermometers gained popularity in both hospital and non-hospital settings due to their ease of use, cleanliness, user-friendliness, and ability to quickly provide massive contactless temperature readings. Additionally, there was a growing dislike of pediatricians and children's parents about invasive devices, such as rectal thermometers [77].

Infrared sensors used in both oximeters and some thermometers register the emissivity of the skin. We see different skin colors due to the effect of skin pigmentation in visible light absorption. Could skin pigmentation also affect infrared thermometer readings? A study with 65 healthy participants showed no significant differences in emissivity between participants when grouped by skin pigmentation (Fitzpatrick scale, p = 0.859, or reflectance spectrophotometry, p = 0.346) [80]. Nevertheless, it is still needed to test medical devices in various environments (such as homes, hospitals, and ICUs) and across more diverse patient populations.

A recent study by Bhavani et al. [25] showed that IR-based thermometers may be prone to calibration discrepancies among darker-pigmented patients. The performance of temporal (IR-based) and oral (contact) thermometers in detecting fever was compared to Black (n = 2031) and White (n = 2344) patients with suspected infection. As opposed to oral measurements, temporal measurements were associated with significantly lower odds of identifying fever among Black patients at multiple fever cutoffs, but not in White patients [25].

These measurement disparities raise questions about whether some thermometry measurements may inadvertently overlook hypothermia, fever or sepsis cases. Disparities of 0.2°C can

drastically change clinical decisions, such as exposing the patient to antimicrobial therapy, drawing blood cultures and isolation [81], which might potentially lead to delayed diagnoses and ultimately exacerbate mortality rates within vulnerable subpopulations.

Considering the disparities verified in some medical devices' measurements and the associated consequences in the patient's healthcare, the solution could be the replacement of the devices with new ones, better calibrated and more reliable. However, for example, it was estimated that the replacement of pulse oximetry hardware in United States hospitals would cost an average of $15,704.12 (€14,644.09) per bed [82]. Approaches focused on software instead of hardware look more viable. One study used a dataset with paired SaO_2 and SpO_2 measurements, patient demographics, physiological data, and treatment information to predict SaO_2 from SpO_2. With XGBoost regression, a $R^2 = 67.6\%$ among Black patients was achieved, mitigating some of the hidden hypoxemia cases [83]. Although it was a single-center study, this seems a more adequate path than medical device replacement.

3.4 Summary

Medical devices are key tools in the healthcare environment. Thus, their design and validation must follow the official recommendations, like the FDA 510(k), and be transparent. Medical devices' regulation must be improved and a continuous evaluation of their accuracy and performance cannot be neglected. Moreover, sharing clinical data could accelerate measurement disparities identification and help mitigate machine learning bias. Pulse oximeters and IR-based thermometers are two examples of devices that might potentially lead to inequities in healthcare diagnoses and treatments, exacerbating mortality rates among vulnerable subpopulations.

By merging this information with Chapter 2, it becomes clear that all steps related to a machine learning framework should be carefully analyzed.

Chapter 4

Methodology

This Chapter outlines the methodology designed to assess and mitigate clinical measurement disparities, focusing on two medical devices: the oximeter and the temporal thermometer. It includes a description of the derived datasets' building process and the statistical and machine learning-based strategies to test this dissertation's hypothesis.

4.1 Datasets

PhysioNet [84] is a platform containing freely available medical data maintained by the Massachusetts Institute of Technology Laboratory for Computational Physiology (MIT-LCP). The project was created to increase data accessibility and promote biomedical research.

BOLD, a blood-gas and oximetry linked dataset [85], was recently published in PhysioNet to facilitate and promote health equity research, targeting bias in pulse oximetry measurements. Similarly, a standardised dataset with paired temperatures and time-aligned patient information was built.

Both datasets used in this study were derived from three PhysioNet databases:

- Medical Information Mart for Intensive Care (MIMIC)-III: contains de-identified health-related data of over 40,000 patients admitted to the ICU of the Beth Israel Deaconess Medical Center (BIDMC); data was collected between 2001 and 2012 [86];
- MIMIC-IV: contains de-identified health-related data of over 70,000 patients admitted to the ICU of the BIDMC between 2008 and 2019 [30];
- eICU Collaborative Research Database (CRD): contains de-identified health-related data of almost 140,000 patients, corresponding to over 200,000 admissions to ICUs of multiple hospitals across the United States between 2014 and 2015 [31].

To access them, the user must have a credentialed account, sign a Data Use Agreement and complete the Collaborative Institutional Training Initiative (CITI) Data or Specimens Only Research course.

4.1.1 Blood-gas and Oximetry Linked Dataset

BOLD [85] is a dataset with paired pulse oximetry readings (SpO_2) and preceding arterial blood gas measurements (SaO_2), acquired within a 5-minute interval. Only values between 70% and 100% are included. Patient characteristics, vital signs, laboratory values, and SOFA scores are time-aligned with the SaO_2 sample. The race and ethnicity information is grouped in the following categories: "American Indian/Alaska Native", "Asian", "Black", "Hispanic OR Latino", "More Than One Race", "Native Hawaiian/Pacific Islander", "Unknown", and "White".

This dataset is restricted to the first pair per patient hospital admission. However, the one used in this study is an extended version, containing all the existing pairs on each hospitalization, under the remaining inclusion criteria.

4.1.2 Thermometry-Linked Dataset

A standardized dataset with paired temperatures and time-aligned patient contextualization information was built. As mentioned previously, disparities in body temperature measurements were verified when using IR-based thermometers instead of contact thermometers [25]. Thus, this dataset might facilitate temperature-related retrospective studies and promote research on racial and ethnic healthcare disparities.

4.1.2.1 Data selection

Reading of reference (contact thermometers: oral, esophageal, and rectal) and IR (temporal and tympanic) temperature values were paired when both measurements occurred within a 4-hour time window, without requiring one to occur before the other. Only values ranging from 30°C to 45°C were considered and missing temperatures or their timestamps and measurement sites were not allowed.

Individual characteristics and time-varying data were aligned with the temperature pairs by the reference temperature measurement timestamp to provide complete information about the patient's situation. Patient admission characteristics, temperature measurement information, vital signs, laboratory values, and SOFA scores were stored.

Uniformity was ensured across all databases by standardizing each variable and only including the ones present in all databases to reduce missing data. As in BOLD, the race and ethnicity original information was mapped to one of the following categories: "American Indian/Alaska Native", "Asian", "Black", "Hispanic OR Latino", "More Than One Race", "Native Hawaiian/Pacific Islander", "Unknown", and "White".

Finally, the standardized databases were vertically concatenated. Each entry has three new unique identifiers:

- *unique_subject_id*, which identifies the patient;
- *unique_hospital_admission_id*, corresponding to a single hospital stay;

- *unique_icustay_id*, which identifies a unique stay in the intensive care unit.

However, the original identifiers (*subject_id*, *hospital_admission_id* and *icustay_id*) were not removed to allow the correspondence to the original databases, which is an advantage in case the user needs more patient's information.

4.1.2.2 Variables description

Variables can be included in one of the following categories: patient admission characteristics, temperature measurement information, vital signs, laboratory values, and SOFA scores. They were all time-aligned with the reference temperature measurement timestamp and the nearest sample was kept. Columns with a *delta* prefix contain information about the time difference between the temperature and variable measurements.

Unlike MIMIC databases, eICU-CRD contains information about several hospitals. Each one has a unique identifier and other hospital-related information. To ensure consistency, MIMIC data was given the hospital index 9999 (does not correspond to any eICU hospital identifier), the number of beds as "≥ 500", the US region as "Northeast", and the teaching status as "True".

Patient and admission characteristics contain information about age, sex, weight, height, body mass index (BMI), race and ethnicity, comorbidities, in-hospital mortality, and length of stay of each hospital and ICU admission. Age information was standardized by considering 90 all ages equal or higher than 90. Patients under 18 years old were excluded. BMI at admission time was calculated with the respective weight and height. MIMIC-III uses the van Walraven Elixhauser score [87] to quantify comorbidities, while MIMIC-IV and eICU-CRD use the Charlson Comorbidity Index [88].

The **temperature information** includes the timestamp, value, and site of a temperature measurement. The columns identified with "1" correspond to the reference temperature and those with "2" to the IR ones. The differences between the reference and the IR-based thermometry values and timestamps are also available.

Vital signs include heart and respiratory rates; systolic, diastolic and mean blood pressure (both invasive and non-invasive); and blood oxygen saturation levels (SpO_2). These variables can be easily identified by the prefix *vitals*. They were pulled from the *chartevents* and *nursecharting* tables of the original MIMIC and eICU databases, respectively.

Laboratory test values are important to characterize each patient's health status. The pulled variables were: Alanine Aminotransferase (alt), Albumin, Alkaline Phosphatase (alp), Anion Gap, Aspartate Aminotransferase (ast), Bicarbonate, Blood Urea Nitrogen (bun), Calcium, Chloride, Creatine Kinase MB (ck_mb), Creatine Phosphokinase (ck_cpk), Creatinine, Fibrinogen, Glucose, Hematocrit, Hemoglobin, International Normalized Ratio (inr), Lactate, Lactate Dehydrogenase (ld_ldh), Mean Corpuscular Hemoglobin (mch), Mean Corpuscular Hemoglobin Concentration (mchc), Mean Corpuscular Volume (mcv), Partial Thromboplastin Time (ptt), Platelet Count, Potassium, Prothrombin Time (pt), Red Blood Cell Count (rbc), Red Cell DistributionWidth (rdw),

Sodium, Total and Direct Bilirubin and White Blood Cell Count (wbc). Data was collected from the *labevents* and *labs* tables of the original MIMIC and eICU databases, respectively.

SOFA (Sequential Organ Failure Assessment) score describes organ dysfunction and allows us to quantify patient morbidity. Coagulation, liver, cardiovascular, central nervous system, renal, and a global score were extracted from the *pivoted_sofa*, *sofa*, and derived tables of MIMIC-III, MIMIC-IV, and eICU databases, respectively. These selected values correspond to the scores one hour before the reference temperature timestamp (identified by the prefix *sofa_past*) or one hour after that time (identified by the prefix *sofa_future*).

4.2 Fever prevalence assessment

As mentioned in Section 3.3, Bhavani et al. [25] found racial differences in fever detection when using temporal instead of oral temperature readings. The study's methodology was reproduced using MIMIC-IV. In order to be able to make a better comparison, more strict than in 4.3.1.2 inclusion criteria were applied to the thermometry dataset: only patients with suspected infection were included; measurements had to occur within a 1-hour window; and only the first temperature pair per ICU admission was kept. Patients had suspected infection if there was a microbiology culture less than or equal to 72h before the administration of the antibiotics or if the antibiotics time was less than or equal to 24h before the culture. Only Black and White patients with oral (contact) and temporal (IR-based) measurements were considered. Paired t-Test was used to compare the temperature readings in each group. Unadjusted and adjusted odds of fever were calculated for the different groups, using oral and temporal thermometry. Unadjusted odds are the odds calculated with the original temperature values. The adjusted odds are calculated with the original oral values and the temporal temperatures obtained with a LR adjusted for age, sex and comorbidities.

Additionally, the unadjusted and adjusted odds of hidden fever for the two subgroups were evaluated. Hidden fever is present when the oral temperature is above the fever threshold but not the temporal temperature. Several fever cutoffs were considered: 37.8°C, 38.0°C, 38.3°C, and 38.5°C. A 95% CI was utilized.

The odds ratio (OR) compares the odds of an event occurring with a specific exposure with the odds of occurring in the absence of that exposure [89]. The following OR were computed:

$$OR\ of\ having\ fever = \frac{Odds\ of\ having\ fever\ using\ the\ temporal\ thermometer}{Odds\ of\ having\ fever\ using\ the\ oral\ thermometer} \tag{4.1}$$

$$OR\ of\ having\ hidden\ fever = \frac{Odds\ of\ having\ hidden\ fever\ in\ Black\ patients}{Odds\ of\ having\ hidden\ fever\ in\ White\ patients} \tag{4.2}$$

4.3 Machine Learning pipeline

The developed machine learning pipeline was applied to both datasets and is now described step by step. The Thermometry-Linked Dataset was restricted to pairs with temperature measurements

occurring within a 1-hour time window. Tympanic temperatures were excluded because they show more variability caused by factors other than the device's technology, as the probe's direction when inserted [77].

In this part of the methodology, hidden hypoxemia was considered as having SaO_2 below 88% and SpO_2 above that threshold (Figure 4.1). Hidden hypothermia and hidden fever were defined as having a temporal (non-contact) measurement in the normal range (36.0-38.0°C) and the reference (contact) temperature below 36.0°C or above 38.0°C, respectively (Figure 4.2).

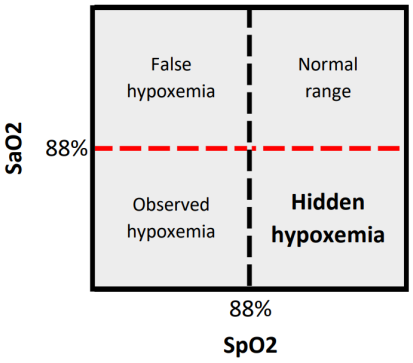


Figure 4.1: Blood oxygen saturation outcomes.

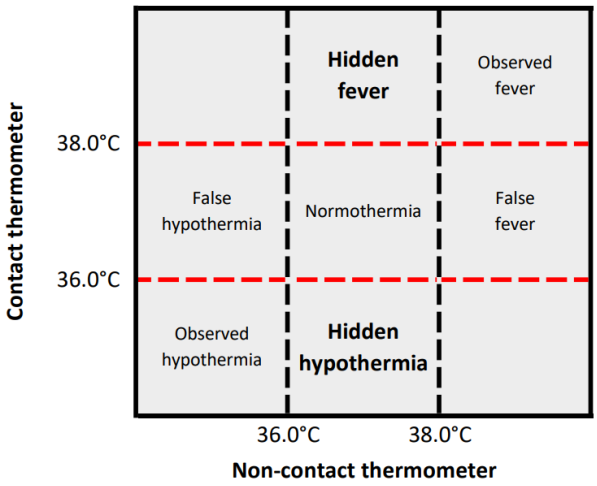


Figure 4.2: Temperature outcomes.

4.3.1 Preprocessing and Feature selection

4.3.1.1 Derived information

It was considered that some relevant information was not explicitly present in the dataset so it was computed as follows:

- Respiratory SOFA score from/in the previous/next 24h: calculated by the difference between the overall SOFA score from/in the previous/next 24h and the sum of the remaining individual scores (coagulation, liver, cardiovascular, central nervous system and renal);
- Binarised respiratory SOFA score in the next 24h: scores above 0 were considered as "1";
- Overall SOFA score increase: class "True" was assigned to cases where the overall SOFA score increased two or more points from the previous to the next 24h after the timestamp, and "False" otherwise.

Due to the low number of samples of patients from race and ethnicity groups "American Indian/Alaska Native", "More Than One Race" and "Native Hawaiian/Pacific Islander", they were grouped with the "Unknown" cases in the class "Other".

4.3.1.2 Manual feature selection

Features with more clinical relevance for the study were manually chosen, based on clinicians' knowledge and the relevance shown in several studies with similar goals (study 1: [4], study 2: [3], study 3: [15]). The selected features are presented in Table 4.1, as well as the indication if they match with the ones considered in the literature.

4.3.1.3 Missing data

When the patient's symptoms do not indicate possible changes in certain laboratory and/or vital values, that information is usually not acquired. Thus, it was replaced with the midpoint of the normal range to a healthy adult [14], which is presented in Table 4.1. The "Temperature" and "SpO_2" were only replaced in the oximetry and thermometry studies, respectively.

Table 4.1: List of the manually selected features for this study and the respective midpoint of the normal range to a healthy adult [14]. It is also shown, with a check mark, which features match the top 15 of Study 1 [4], the top 17 of Study 2 [3] and the top 10 of Study 3 [15].

Variable	Midpoint	Units	Study 1	Study 2	Study 3
Age	-	years	✔	✔	
Albumin	4.4	g/dL			
Anion Gap	10.0	mEq/L	✔	✔	
Bicarbonate	25.0	mEq/L	✔	✔	
Blood Urea Nitrogen	13.5	mg/dL		✔	✔
Comorbidity score	-	-			
Creatinine	0.9	mg/dL			
Glucose	84.5	mg/dL			✔
Heart Rate	80	bpm		✔	✔
Hemoglobin (Men)	15.5	g/dL		✔	
Hemoglobin (Women)	13.75	g/dL		✔	
Lactate	01.05	mmol/L		✔	
Mean Blood Pressure	87.5	mmHg		✔	
Platelet Count	300.0	$x10^3$/uL	✔	✔	✔
Potassium	4.25	mEq/L		✔	
Red Blood Cell Count	5.0	$x10^6$/uL			
Red Cell Distribution Width	13.0	%			✔
Respiratory Rate	16	breaths/min	✔	✔	✔
Sex	-	-		✔	
Sodium	140	mEq/L	✔	✔	
SOFA score (past 24h) - Cardiovascular	-	-			
SOFA score (past 24h) - Overall	-	-			
SOFA score (past 24h) - Respiratory	-	-			
SpO_2	95	%		✔	✔
Temperature	36.65	°C	✔	✔	✔
White Blood Cell Count	7.25	$x10^3$/uL	✔	✔	✔

4.3.2 Machine Learning models

The following binary classification tasks were chosen: in-hospital mortality, respiratory SOFA score in the next 24h and overall SOFA score increase. Patients without information about in-hospital mortality or at least one SOFA score component were excluded. The dataset was divided into train and test sets with the Stratified K-Fold cross-validator, using 10 folds. Considering the dataset and the classification problems, three models were tested: Logistic Regression, Random Forest and XGBoost. Class weights were applied to the different classes because of data imbalance. The following performance metrics were computed: AUROC, recall, F1-score and accuracy. Demographic parity difference and Equalized odds ratio were additionally calculated, considering the race and ethnicity group as the sensitive feature.

4.3.3 Bias mitigation

As an attempt to correct the possibly biased feature, three strategies were designed:

1. The difference between the reference value and the biased feature of the first pair in each hospital admission was calculated. Then, that value was added to the biased feature of the following pairs from the same hospital admission.

2. XGBoost Regressor was used to predict the reference feature based on the remaining patient information (including the biased feature). This way, the model could learn how to correct the biased feature and then be applied in the test set. This strategy only used the first pair per hospital admission.

3. The third approach was the same as the second, but the model was trained with all pairs.

To evaluate these models' performance, RSME and R^2 were calculated, considering the SaO_2 and reference temperature as true values and the SpO_2 and temporal temperature (corrected or not) as predictions, in the respective studies.

This part of the methodology was built under counterfactual thinking, although no counterfactual models were used. The objective of the pipeline was to compare the models' performance by changing a single feature and observing the differences in the outcomes. Thus, there were five scenarios, each considering one of the following features: reference feature, biased feature and biased feature corrected with strategies 1, 2 or 3. Paired t-Test was used to compare the model with the reference feature with the remaining ones.

Figure 4.3 summarises the machine learning framework designed in this dissertation. As previously explained, regarding blood oxygen saturation, SaO_2 is the reference value and SpO_2 is the biased one. In thermometry, oral/core and temporal are the reference and biased features, respectively. The "Biased feature correction" step is in a dashed box because it is not applied when the original SpO_2 or temporal values are compared with the reference values.

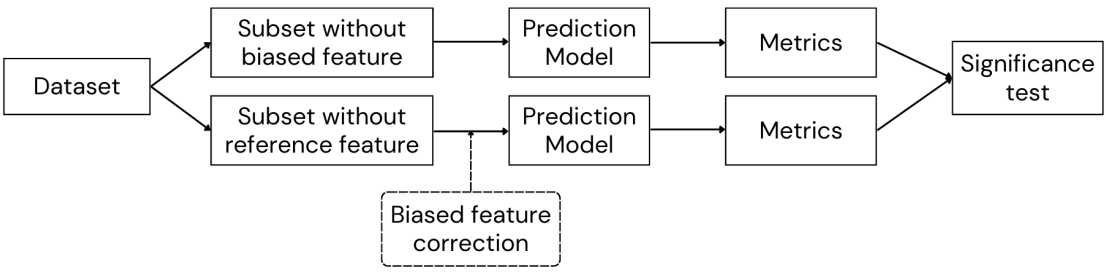


Figure 4.3: Machine Learning pipeline for the identification and mitigation of clinical measurement disparities.

Chapter 5

Results and discussion

The present Chapter outlines the selected cohort of the two derived datasets, the results of the fever prevalence assessment in the thermometry dataset and the performance of the ML pipeline.

5.1 Datasets

5.1.1 Blood-gas and Oximetry

The extended version of the BOLD dataset includes a total of 203,087 pairs, representing 44,902 patients. Out of these pairs, 8,964 pairs (4,774 patients) were sourced from MIMIC-IV; 1,273 pairs (727 patients) from MIMIC-III; and 192,850 pairs (39,401 patients) from eICU-CRD. Table 5.1 briefly describes the patients' personal and clinical information in the selected cohort. Figure 5.1a shows the percentage of patients per sex and race and ethnicity group.

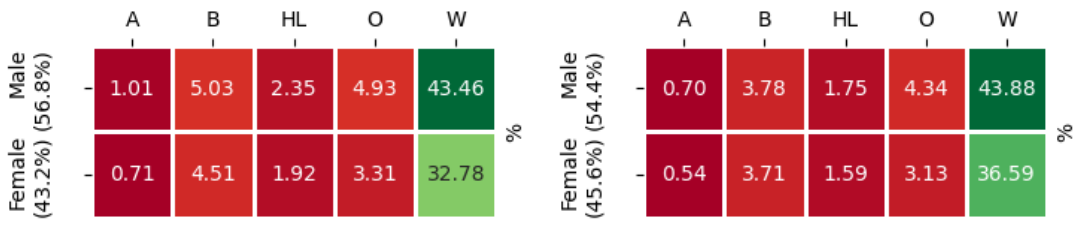
5.1.2 Thermometry

The thermometry-linked dataset includes a total of 83,633 pairs, representing 20,787 patients. 21,277 pairs (4,099 patients) were sourced from MIMIC-IV; 114 pairs (44 patients) from MIMIC-III; and 62,242 pairs (16,644 patients) from eICU-CRD. Table 5.2 presents patients' personal and clinical information that is relevant to characterise the selected cohort. Figure 5.1b discriminates the percentage of patients per sex and race and ethnicity group. Additionally, Figures A.1, A.2 and A.3 from Appendix A show the flow diagrams of the process applied to each original database that allowed to achieve the final dataset.

Both datasets have more pairs than patients as a single patient can have multiple hospital stays. The percentage of patients per race and ethnicity groups is unbalanced and there is a slightly higher percentage of male subjects in each group.

Table 5.1: Patient information from the cohort for the blood-gas and oximetry study.

	eICU-CRD	**MIMIC-III**	**MIMIC-IV**
n	39401	727	4774
Age, median [IQR]	66.0 [55.0,76.0]	68.0 [58.0,77.0]	68.0 [59.0,77.0]
Sex Female, n (%)	17448 (44.3)	278 (38.2)	1683 (35.3)
In-Hospital Mortality, n (%)	7257 (18.4)	129 (17.7)	742 (15.5)
Race/Ethnicity, n (%)			
- Asian	644 (1.6)	16 (2.2)	112 (2.3)
- Black	3915 (9.9)	41 (5.6)	325 (6.8)
- Hispanic or Latino	1740 (4.4)	20 (2.8)	156 (3.3)
- Other	2770 (7.0)	72 (9.9)	859 (18.0)
- White	30332 (77.0)	578 (79.5)	3322 (69.6)
Weight, median [IQR]	81.4 [67.1,99.0]	82.4 [70.0,97.5]	82.0 [70.0,97.0]
Height, median [IQR]	170.1 [162.5,177.8]	170.2 [162.6,177.8]	170.0 [163.0,178.0]
BMI, median [IQR]	28.0 [23.7,33.5]	28.3 [24.7,33.4]	28.3 [24.7,32.7]
Hidden Hypoxemia, n (%)	1294 (3.3)	15 (2.1)	23 (0.5)
Hospital LoS (dead), median [IQR]	5.5 [2.3,11.6]	11.0 [5.0,23.0]	12.0 [6.0,21.0]
Hospital LoS (alive), median [IQR]	8.3 [5.2,14.0]	10.0 [7.0,18.0]	10.0 [6.0,18.0]
ICU LoS (dead), median [IQR]	3.2 [1.4,6.8]	9.0 [3.0,17.0]	8.9 [4.1,15.0]
ICU LoS (alive), median [IQR]	2.9 [1.7,5.7]	4.0 [2.0,10.0]	3.8 [2.0,8.6]
Comorbidity Score, median [IQR]	4.0 [2.0,6.0]	9.0 [3.0,16.0]	5.0 [3.0,7.0]
N pairs (per hosp. adm.), median [IQR]	2.0 [1.0,5.0]	1.0 [1.0,2.0]	1.0 [1.0,2.0]
SOFA Past Overall 24hr, median [IQR]	5.0 [3.0,8.0]	8.0 [4.0,10.0]	6.0 [4.0,8.0]
SOFA Future Overall 24hr, median [IQR]	5.0 [3.0,7.0]	7.0 [5.0,10.8]	6.0 [4.0,8.0]



(a) Blood-gas and Oximetry Dataset.

(b) Thermometry Dataset.

Figure 5.1: Patients distribution per sex and race/ethnicity group (A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White).

Table 5.2: Patient information from the cohort for the thermometry study.

	eICU-CRD	**MIMIC-III**	**MIMIC-IV**
n	16644	44	4099
Age, median [IQR]	65.0 [53.0,76.0]	63.5 [54.0,76.2]	70.0 [59.0,80.0]
Sex (Female), n (%)	7658 (46.0)	21 (47.7)	1790 (43.7)
In-Hospital Mortality, n (%)	911 (5.5)	6 (13.6)	1000 (24.4)
Race/Ethnicity, n (%)			
- Asian	135 (0.8)	2 (4.5)	121 (3.0)
- Black	1117 (6.7)	2 (4.5)	437 (10.7)
- Hispanic or Latino	558 (3.4)	1 (2.3)	135 (3.3)
- Other	776 (4.7)	5 (11.4)	771 (18.8)
- White	14058 (84.5)	34 (77.3)	2635 (64.3)
Weight, median [IQR]	81.0 [67.3,97.5]	68.7 [62.6,88.5]	76.0 [64.0,91.0]
Height, median [IQR]	170.2 [162.6,177.8]	170.2 [162.6,177.8]	170.0 [163.0,178.0]
BMI, median [IQR]	27.8 [23.8,33.0]	24.9 [22.5,28.5]	27.1 [23.3,31.9]
Hidden Fever, n (%)	441 (2.6)	3 (6.8)	118 (2.9)
Hidden Hypothermia, n (%)	350 (2.1)	6 (13.6)	62 (1.5)
Hospital LoS (dead), median [IQR]	7.9 [3.6,15.7]	14.0 [5.8,16.2]	12.0 [6.0,21.0]
Hospital LoS (alive), median [IQR]	6.1 [3.3,11.0]	7.0 [5.0,18.8]	11.0 [7.0,22.0]
ICU LoS (dead), median [IQR]	3.0 [1.2,6.4]	6.0 [3.5,12.2]	6.4 [3.0,11.9]
ICU LoS (alive), median [IQR]	1.8 [1.0,3.2]	3.0 [1.0,10.5]	3.5 [1.9,7.8]
Comorbidity Score, median [IQR]	3.0 [2.0,5.0]	7.0 [0.0,13.0]	6.0 [4.0,8.0]
N pairs (per hosp. adm.), median [IQR]	2.0 [1.0,4.0]	2.0 [1.0,3.0]	2.0 [1.0,5.0]
Delta temperature (ºC), median [IQR]	0.1 [-0.2,0.5]	-0.2 [-0.6,0.1]	0.1 [-0.2,0.5]
Delta time (hours), median [IQR]	2.9 [1.8,3.6]	4.0 [3.8,4.0]	4.0 [3.0,4.0]
SOFA Past Overall 24hr, median [IQR]	3.0 [1.0,5.0]	4.0 [2.0,8.0]	5.0 [3.0,7.0]
SOFA Future Overall 24hr, median [IQR]	4.0 [2.0,6.0]	7.0 [3.0,10.0]	5.0 [3.0,7.0]

5.2 Fever prevalence

Table 5.3 presents information about the selected cohort. A total of 530 temperature pairs, from 508 patients (13% Black), were included. Fever prevalence in Black patients was lower when using temporal instead of oral thermometers, for thresholds 38.0ºC, 38.3°C, and 38.5°C. For example, for the threshold 38.3°C, the temporal thermometer did not detect almost 6% of the fevers. On the other side, in White patients, fever prevalence was higher when using temporal instead of oral thermometers, for all thresholds (Figure 5.2). This means that there were several false fever identifications.

Temporal measurements were associated with significantly higher OR of fever in White subjects for the threshold 37.8ºC, compared to oral measurements. None of the thresholds result in significant OR of fever in Black patients. Regarding OR of hidden fever, none of the thresholds showed significant results. However, Figure 5.3 allows us to see that OR are consistently in opposite directions, comparing Black and White groups.

Table 5.3: Patient information from the cohort for the fever prevalence study.

	Black	**White**
n	67	463
Age, median [IQR]	66.0 [55.0,77.0]	67.0 [56.0,76.0]
Sex (Female), n (%)	35 (52.2)	196 (42.3)
In-Hospital Mortality, n (%)	20 (29.9)	142 (30.7)
Oral temperature, mean (SD)	36.9 (1.1)	37.0 (0.8)
Temporal temperature, mean (SD)	36.8 (0.9)	37.0 (1.0)
Comorbidity Score, median [IQR]	6.5 (3.2)	5.9 (3.0)

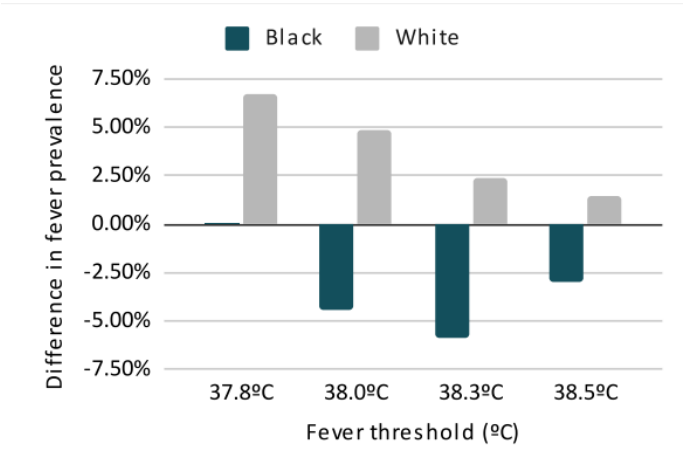


Figure 5.2: Difference in fever prevalence percentage between temporal and oral measurements.

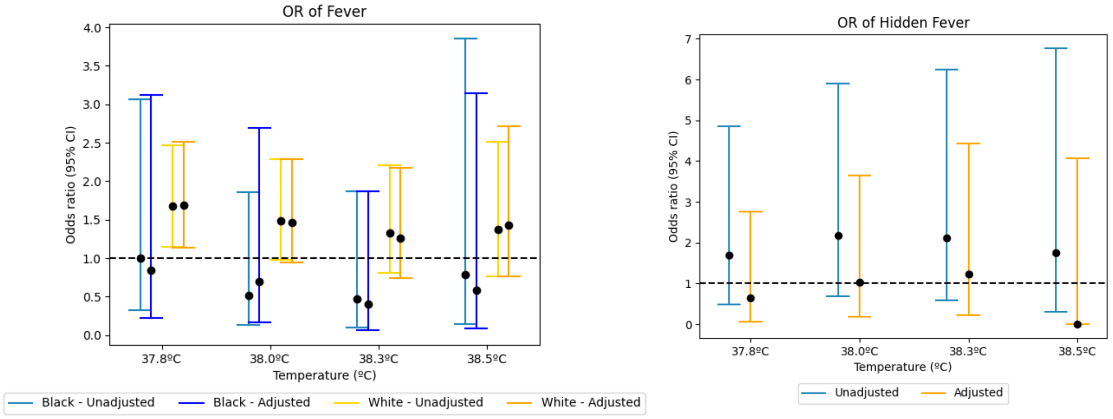
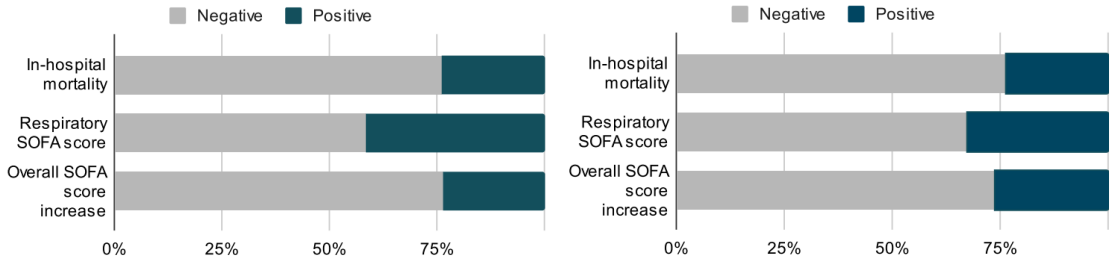


Figure 5.3: OR of having fever (left), comparing temporal with oral measurements, and hidden fever (right), comparing Black with White patients.

5.3 Disparities assessment

The blood-gas and oximetry dataset was reduced to 163,396 pairs from 34,252 patients, after removing subjects without information about in-hospital mortality or at least one SOFA score component (Table 5.4). The thermometry-linked dataset was also under additional exclusion criteria, decreasing the number of pairs and subjects in the cohort to 2,662 and 1,328, respectively (Table 5.5).

The distribution of the observations by the positive and negative classes of each classification task is represented in Figure 5.4.



(a) Blood-gas and Oximetry Dataset.

(b) Thermometry Dataset.

Figure 5.4: Distribution of the observations by the positive and negative classes of each classification task.

Several authors have verified disparities in medical device measurements across subpopulations. In this study, three approaches were considered to evaluate the models' results using the reference or the biased values:

- Across race and ethnicity subgroups (Asian, Black, Hispanic or Latino, Other and White);
- By level and direction of disparity between the two measurements;

Table 5.4: Patient information from the cohort for the disparities assessment on the blood-gas and oximetry study.

	Asian	Black	Hispanic or Latino	Other	White
n	605	3,397	1,448	2,823	25,979
Age, median [IQR]	66.0 [54.0,78.0]	61.0 [51.0,71.0]	67.5 [53.0,78.0]	65.0 [52.0,75.0]	67.0 [57.0,77.0]
Sex Female, n (%)	253 (41.8)	1613 (47.5)	661 (45.6)	1142 (40.5)	11251 (43.3)
In-Hospital Mortality, n (%)	115 (19.0)	589 (17.3)	280 (19.3)	501 (17.7)	4528 (17.4)
Hidden Hypoxemia, n (%)	15 (2.5)	129 (3.8)	40 (2.8)	71 (2.5)	722 (2.8)
Hospital LoS (dead), median [IQR]	9.1 [5.0,17.3]	10.0 [5.0,17.6]	9.1 [4.5,17.2]	8.7 [4.3,17.0]	7.8 [4.0,14.1]
Hospital LoS (alive), median [IQR]	10.1 [6.0,17.9]	10.8 [6.7,18.1]	11.2 [7.1,18.5]	10.1 [6.5,17.0]	9.7 [6.1,16.0]
ICU LoS (dead), median [IQR]	6.2 [3.5,12.3]	6.5 [3.4,11.1]	5.0 [2.8,9.2]	6.0 [3.4,10.7]	5.0 [2.9,9.2]
ICU LoS (alive), median [IQR]	4.0 [2.3,7.8]	4.2 [2.6,7.8]	3.7 [2.2,7.0]	4.1 [2.4,7.9]	3.9 [2.3,7.1]
Comorbidity Score, median [IQR]	4.0 [2.0,6.0]	4.0 [2.0,6.0]	4.0 [2.0,6.0]	4.0 [2.0,6.0]	4.0 [2.0,6.0]
N pairs (per hosp. adm.), median [IQR]	2.0 [1.0,5.0]	2.0 [1.0,5.0]	2.0 [1.0,6.0]	2.0 [1.0,5.0]	2.0 [1.0,5.0]
SOFA Past Overall 24hr, median [IQR]	5.0 [2.0,8.0]	5.0 [3.0,8.0]	5.0 [3.0,8.0]	6.0 [3.0,8.0]	5.0 [3.0,8.0]
SOFA Future Overall 24hr, median [IQR]	4.0 [2.0,7.0]	5.0 [3.0,8.0]	5.0 [3.0,7.0]	5.0 [3.0,8.0]	5.0 [3.0,7.0]

- By distinguishing patients with consistent measurement values within the normal range from the ones with hidden hypoxemia, hypothermia or fever.

5.3.1 SaO_2 vs. SpO_2

Logistic regression results were generally similar in the training and test set, leaving room for performance improvement. In contrast, the Random Forest model overfitted the training data, compromising the test results. The XGBoost metrics dropped slightly on the test set.

Looking at the results across racial and ethnic groups, several situations were identified as significantly different by the LR model (Figure 5.5), mostly in the White patients subgroup and in the overall performance. With XGBoost, significant differences were more distributed across the various subgroups (Figure 5.6). Performance metrics were consistently higher in the SaO_2 model.

Table 5.5: Patient information from the cohort for the disparities assessment on the thermometry study.

	Asian	Black	Hispanic or Latino	Other	White
n	37	116	42	178	955
Age, median [IQR]	67.0 [57.0,76.0]	63.0 [52.0,74.0]	59.5 [46.2,68.8]	62.0 [47.2,73.8]	67.0 [56.0,76.0]
Sex Female, n (%)	19 (51.4)	62 (53.4)	18 (42.9)	70 (39.3)	409 (42.8)
In-Hospital Mortality, n (%)	9 (24.3)	22 (19.0)	13 (31.0)	62 (34.8)	171 (17.9)
Hidden Fever, n (%)	0 (0.0)	8 (6.9)	4 (9.5)	15 (8.4)	37 (3.9)
Hidden Hypothermia, n (%)	4 (10.8)	9 (7.8)	2 (4.8)	5 (2.8)	23 (2.4)
Hospital LoS (dead), median [IQR]	16.6 [12.0,39.0]	16.0 [10.2,40.2]	12.0 [9.0,18.0]	15.0 [4.2,23.0]	14.4 [7.3,25.0]
Hospital LoS (alive), median [IQR]	22.0 [11.8,41.5]	13.1 [6.5,33.1]	16.0 [8.0,34.0]	16.9 [8.5,34.0]	11.0 [6.3,21.0]
ICU LoS (dead), median [IQR]	7.5 [3.7,12.6]	9.7 [6.7,14.4]	9.6 [7.4,13.8]	7.7 [3.8,16.0]	9.2 [4.7,16.0]
ICU LoS (alive), median [IQR]	6.3 [3.4,18.7]	5.2 [2.8,12.9]	4.0 [2.6,20.8]	7.3 [3.6,16.7]	4.5 [2.7,9.4]
Comorbidity Score, median [IQR]	4.0 [3.0,7.0]	5.0 [3.0,8.0]	5.0 [2.0,7.0]	4.0 [2.0,6.8]	4.0 [2.0,6.0]
N pairs (per hosp. adm.), median [IQR]	2.0 [1.0,4.0]	1.0 [1.0,2.0]	1.0 [1.0,2.8]	1.0 [1.0,2.0]	1.0 [1.0,2.0]
Delta temperature (ºC), median [IQR]	-0.1 [-0.8,0.2]	0.0 [-0.3,0.5]	-0.0 [-0.4,0.6]	0.0 [-0.4,0.4]	0.0 [-0.3,0.4]
Delta time (hours), median [IQR]	1.0 [1.0,1.0]	1.0 [0.8,1.0]	1.0 [1.0,1.0]	1.0 [1.0,1.0]	1.0 [0.6,1.0]
SOFA Past Overall 24hr, median [IQR]	6.0 [3.0,9.0]	4.0 [2.0,8.0]	6.0 [3.2,8.0]	5.0 [3.0,10.0]	4.0 [2.0,7.0]
SOFA Future Overall 24hr, median [IQR]	5.0 [3.0,9.0]	5.0 [3.0,8.0]	5.0 [3.0,8.0]	6.0 [3.0,10.0]	4.0 [2.0,7.0]

Random Forest showed great AUROC and accuracy but recall and F1-score were very low for two out of three prediction tasks (Figure 5.7), which can be explained by the model's overfitting (Table B.1 from Appendix B). It correctly identifies the samples in the negative class, which are the majority, and the ones in the positive class are neglected. Due to this issue, the remaining results obtained with RF will not be explored in the discussion, but they are presented in Appendix B.

In the second approach, disparities were divided into four groups, according to the difference between SpO_2 and SaO_2 values: $<$ -3%, between -3% and 0%, between 0% and 3%, and $\geq$ 3%.

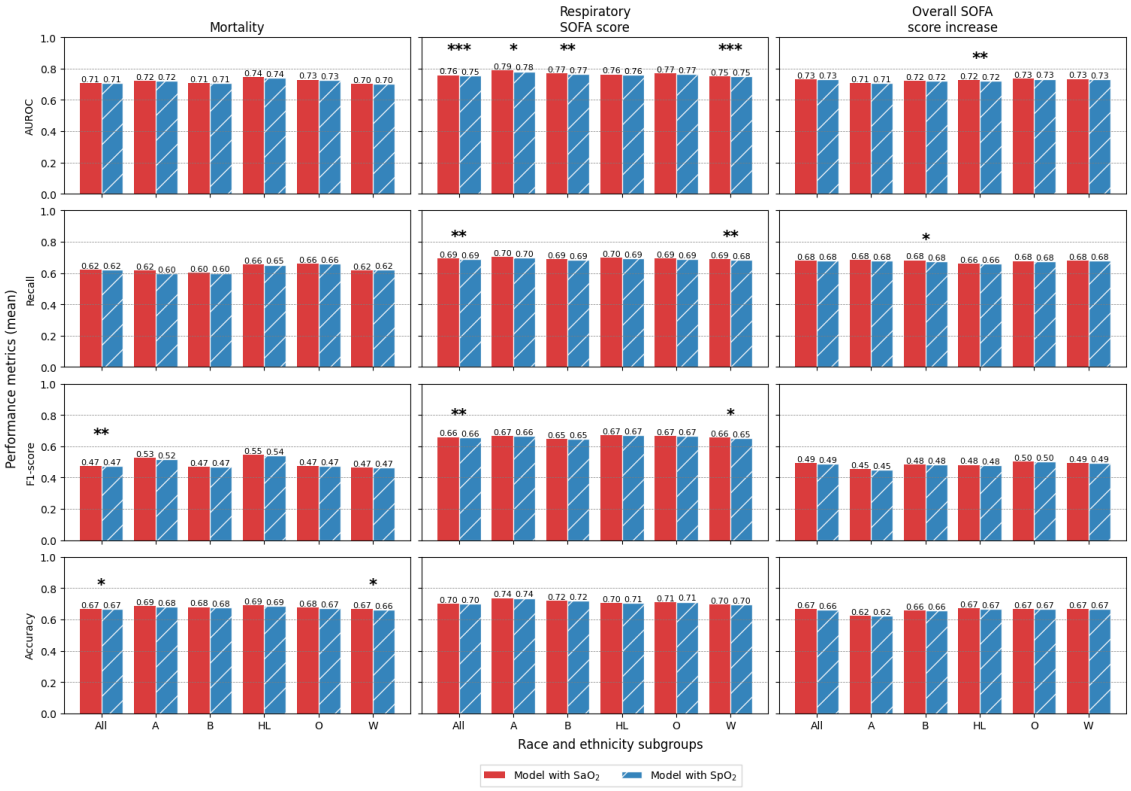


Figure 5.5: Mean values of the LR performance metrics across race and ethnicity subgroups, in the Blood-gas and Oximetry study. Significant differences between SaO_2 and SpO_2 models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001. A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White.

Figures 5.8a and 5.8b represent the bias effect on LR and XGBoost prediction performance, respectively. Figure 5.9 shows the differences in the LR performance metrics between the SpO_2 and SaO_2 models. Accuracy is significantly higher in the SaO_2 model when SpO_2 is underestimated and vice-versa. On the other side, recall is significantly higher in the SaO_2 model when SpO_2 is overestimated and vice-versa. Differences in performance are exacerbated by higher disparities.

The positive class of the three prediction tasks corresponds to an outcome with negative consequences for the patient's health, which means that the negative class might correspond to a less sick population. Thus, the SpO_2 overestimation will possibly increase the correct prediction of the negative cases. In this situation, where the negative class is in the majority, that is the biggest contribution to the accuracy increase. However, the positive class might not be correctly predicted, decreasing the true positive rate (recall).

The last comparison was related to hidden hypoxemia cases. Results are similar to the ones in the previous scenario: accuracy is significantly lower and recall is significantly higher to the SaO_2 model when SpO_2 is overestimated and above or equal to 88% in situations of hypoxemia, meaning that there is a hidden hypoxemia.

Demographic parity difference and Equalized odds ratio were computed considering the race and ethnicity groups as the sensitive features. Low values from the former and high from the latter

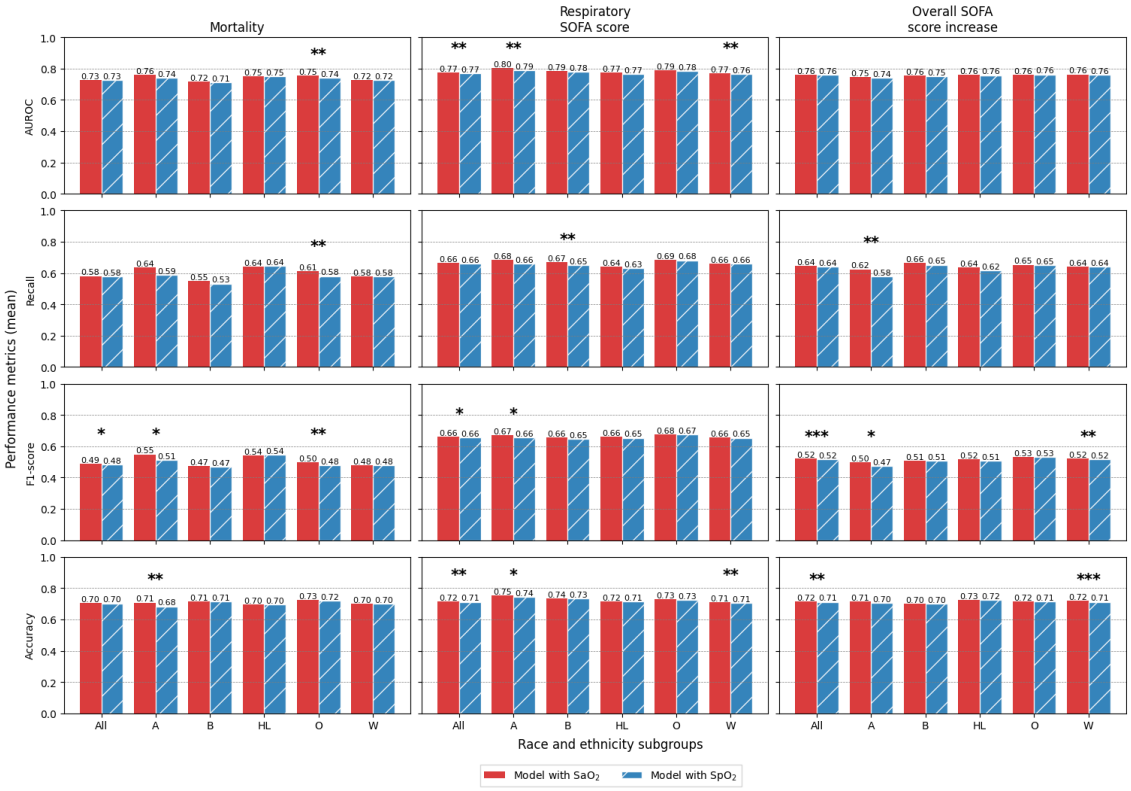


Figure 5.6: Mean values of the XGBoost performance metrics across race and ethnicity subgroups, in the Blood-gas and Oximetry study. Significant differences between SaO_2 and SpO_2 models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001. A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White.

correspond to more fair models. The presented results are referents to a global analysis of LR and XGBoost models and the three prediction tasks. There were no significant differences between the SaO_2 and SpO_2 models. The maximum value of Demographic parity difference was 0.13, meaning that the Predicted Positive Rate was similar to all groups. The minimum Equalized odds ratio was 0.63.

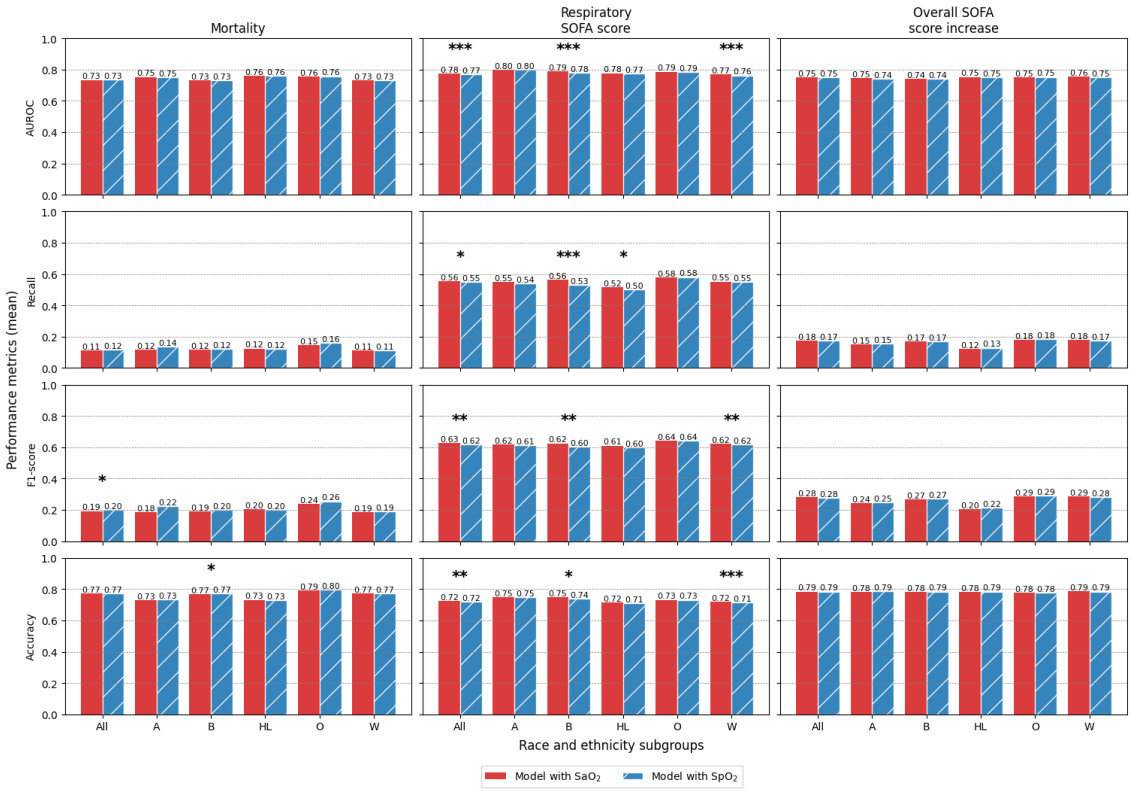
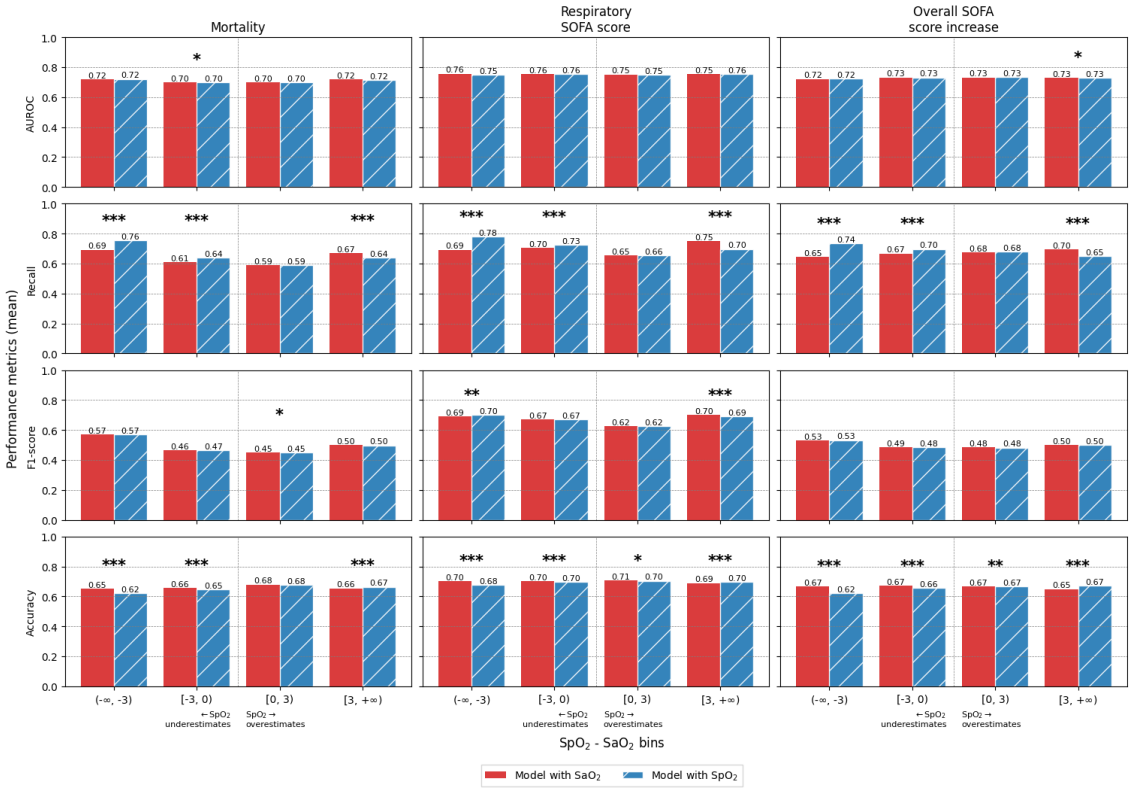
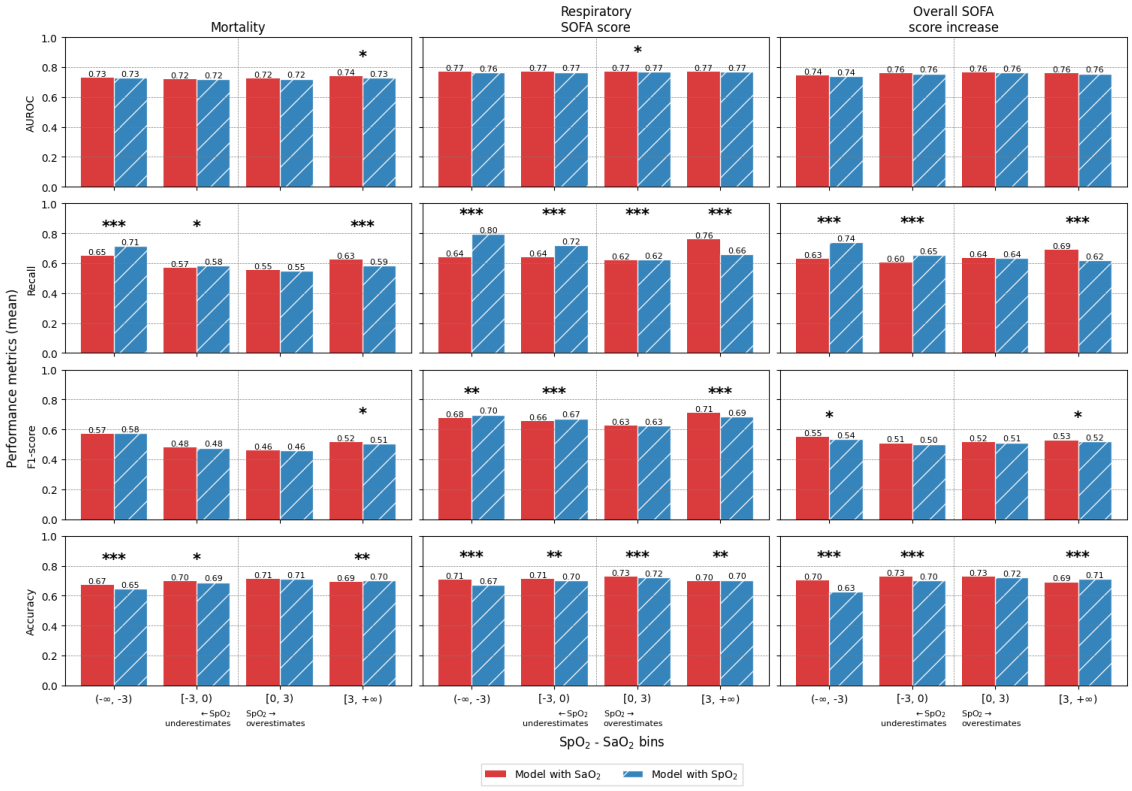


Figure 5.7: Mean values of the RF performance metrics across race and ethnicity subgroups, in the Blood-gas and Oximetry study. Significant differences between SaO_2 and SpO_2 models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001. A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White.



(a) LR



(b) XGBoost

Figure 5.8: Mean value of the performance metrics across disparity groups, in the Blood-gas and Oximetry study. Significant differences between SaO_2 and SpO_2 models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001.

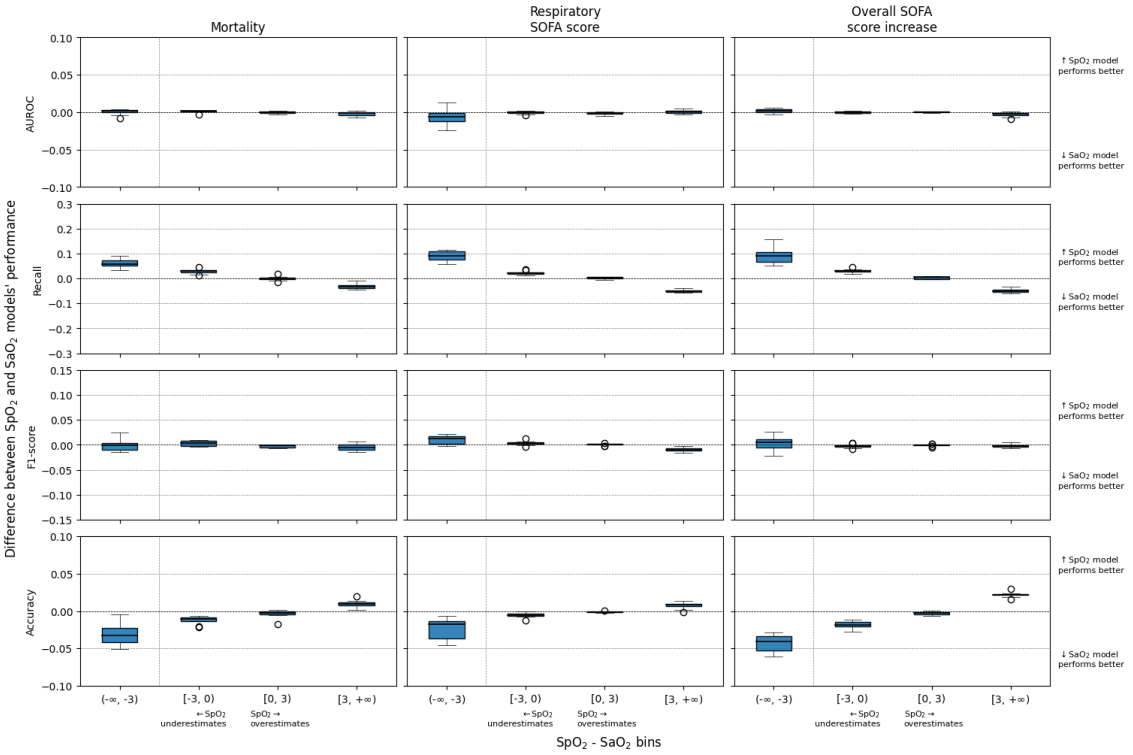
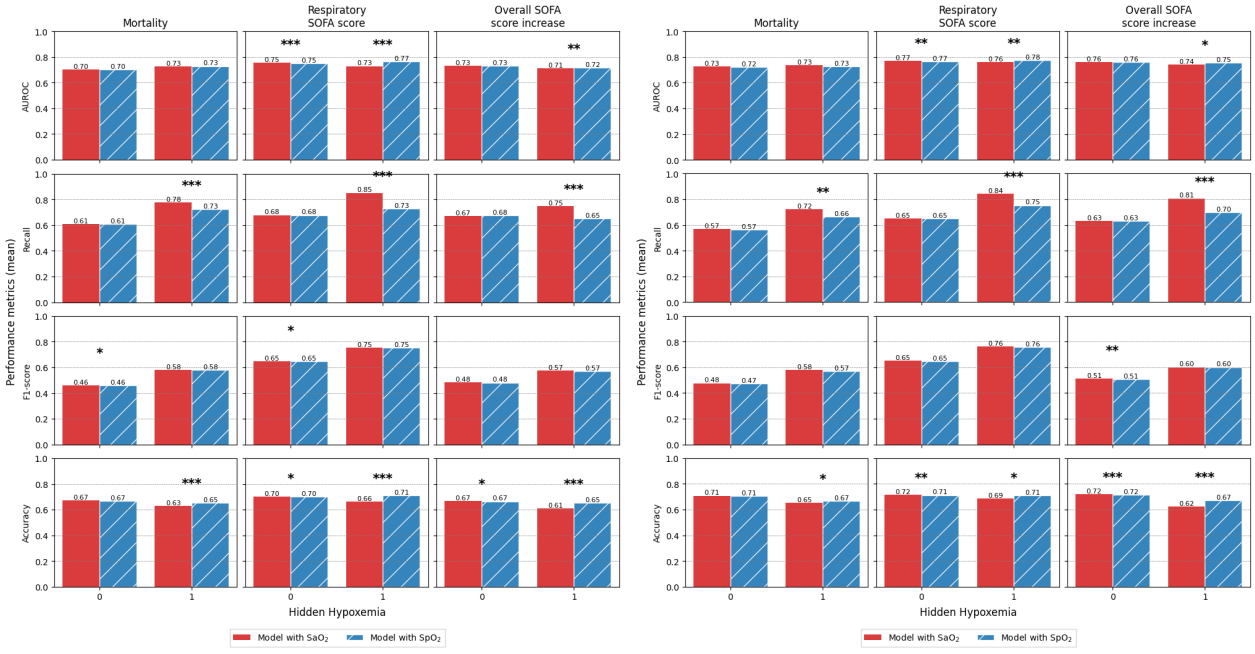


Figure 5.9: Difference in LR performance metrics between the SpO_2 and SaO_2 models, across disparity groups, in the Blood-gas and Oximetry study.



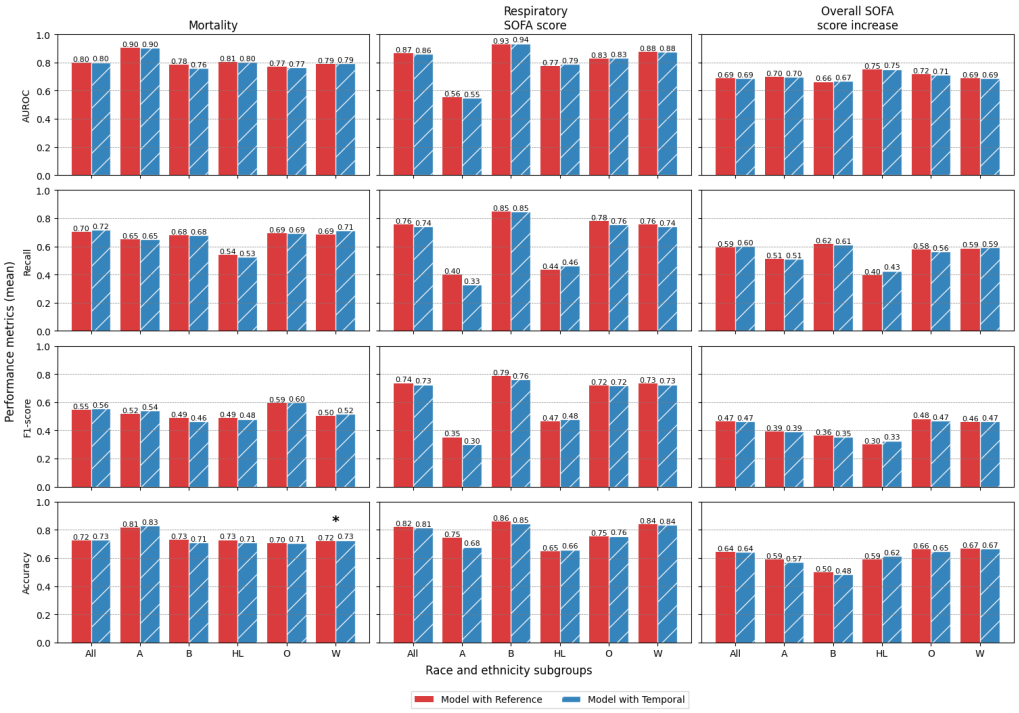
(a) LR (b) XGBoost

Figure 5.10: Mean value of the performance metrics between patients with consistent SaO_2 and SpO_2 values (above or equal to 88%) and the ones with hidden hypoxemia. Significant differences between SaO_2 and SpO_2 models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001.

5.3.2 Contact vs. Temporal

From a global perspective, Logistic Regression's AUROC and accuracy values have stabilized, but recall and F1-score dropped in some cases, from the training to the test set. A similar but more expressive effect was verified in XGBoost, derived from some overfitting (Table B.18 from Appendix B). As previously, RF was overfitted (Table B.18 from Appendix B).

The thermometry dataset was evaluated with the three approaches described in 5.3. Nearly all results across race and ethnicity subgroups were not significantly different between reference and temporal models, using LR, RF or XGBoost (Figures 5.11a, 5.11b and 5.11c, respectively).



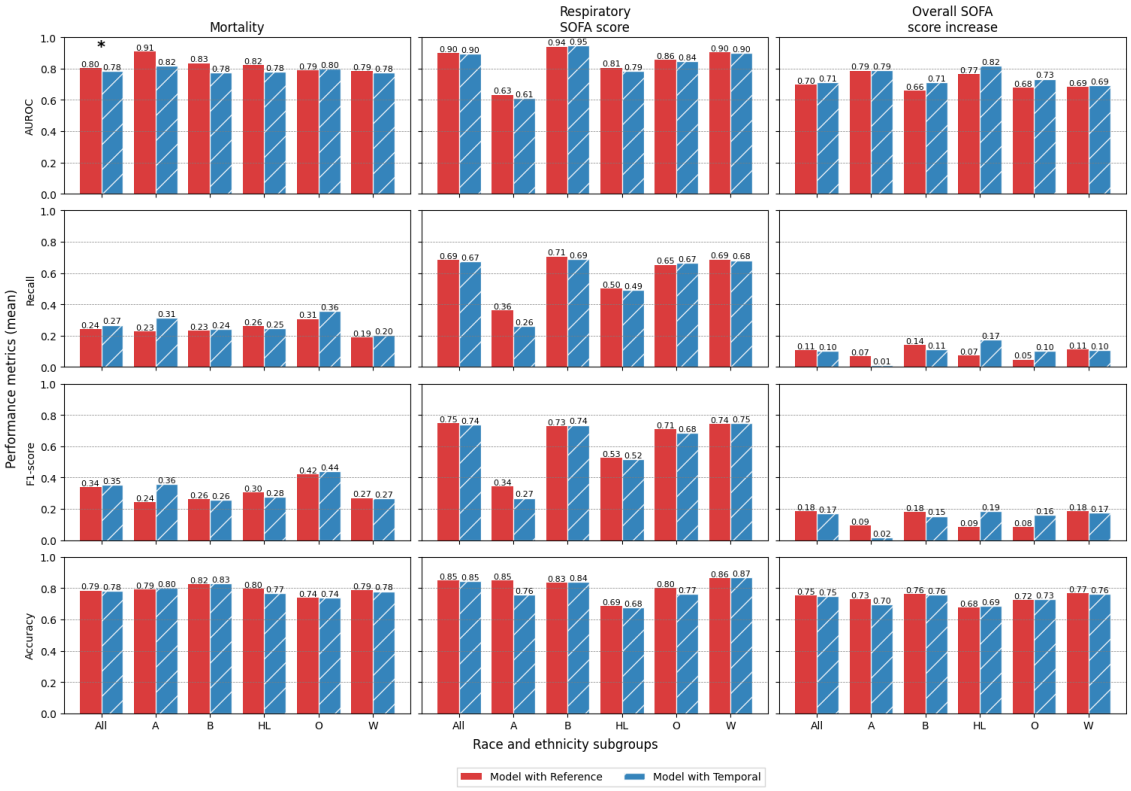
(a) LR

Figure 5.11: Mean value of the performance metrics across race and ethnicity subgroups, in the Thermometry study. Significant differences between reference and temporal models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001. A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White.

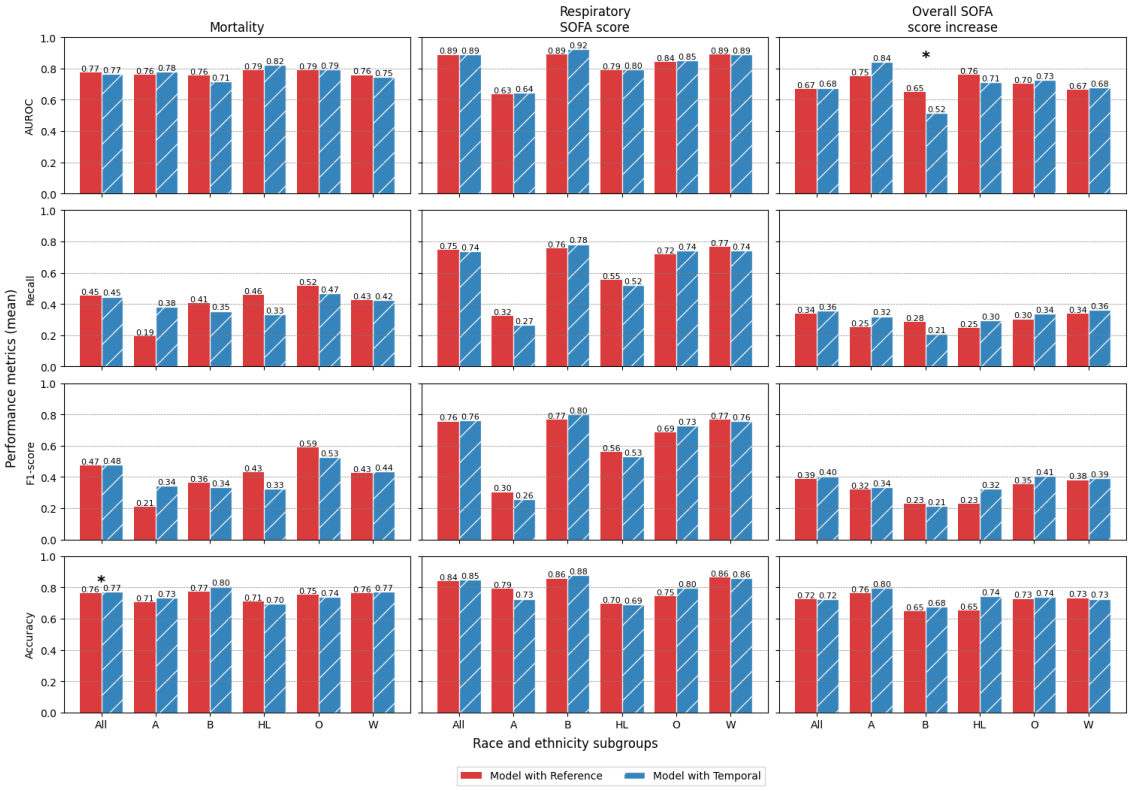
In the second approach, according to the difference between temporal and the reference values, the considered disparity groups were: $<$ -1%, between -1% and 0%, between 0% and 1%, and $\geq$ 1%. The reference and temporal models performed similarly (Figures 5.12a and 5.12b).

In the last approach, significant differences in hidden hypothermia, normothermia and hidden fever groups were nonexistent (Figures 5.13a and 5.13b).

The maximum Demographic parity difference and the minimum Equalized odds ratio were 0.56 and 0, respectively. These values represent considerable differences in the Predicted Positive Rate across groups and distinct confusion matrices among groups.

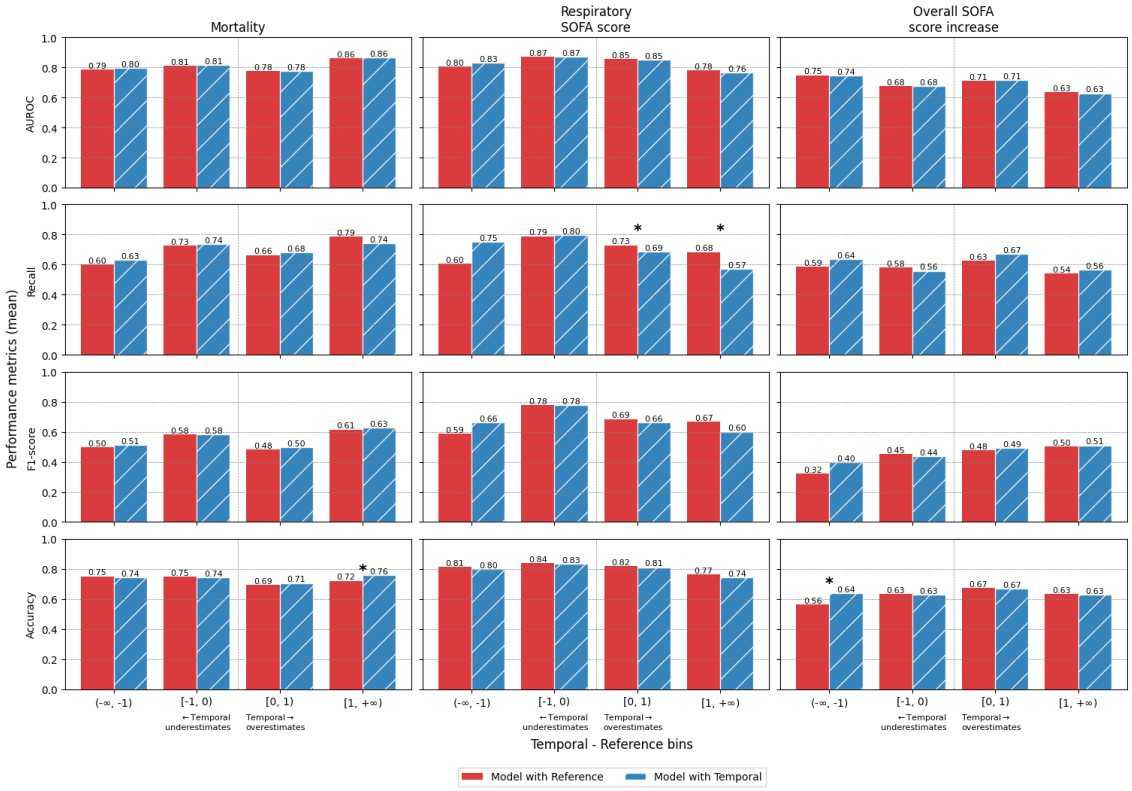


(b) RF

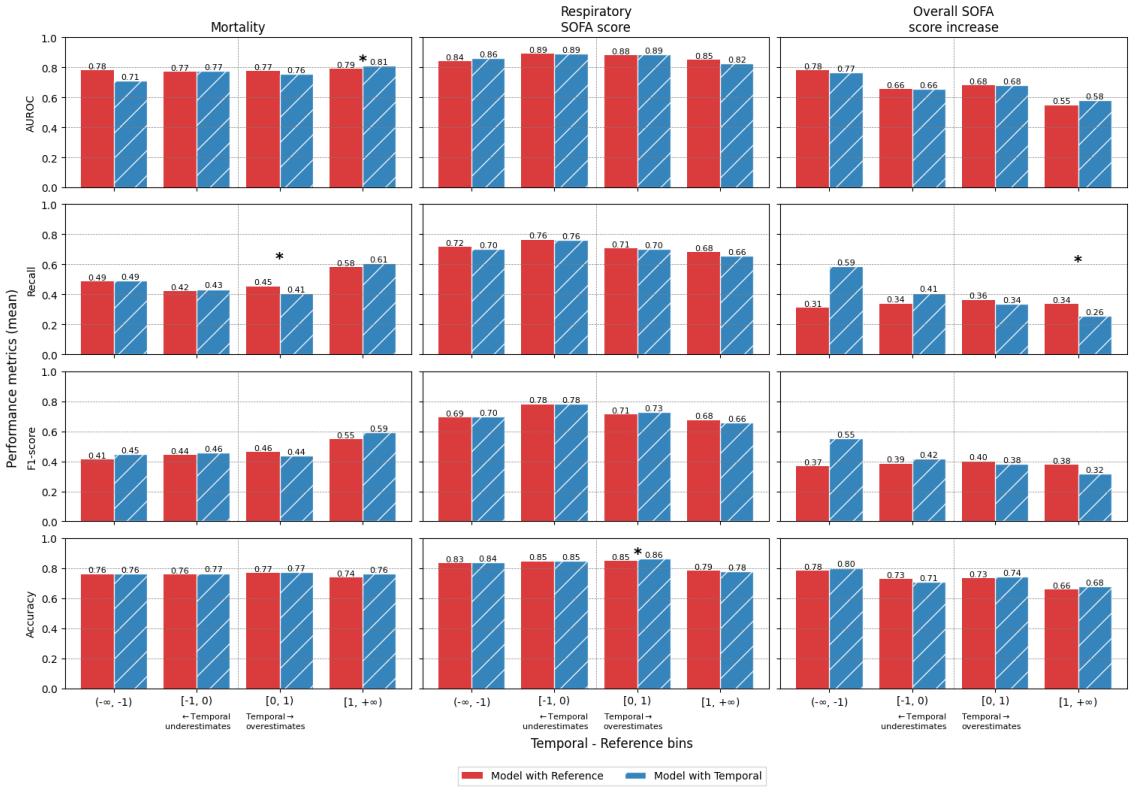


(c) XGBoost

Figure 5.11: Mean value of the performance metrics across race and ethnicity subgroups, in the Thermometry study. Significant differences between reference and temporal models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001. A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White.

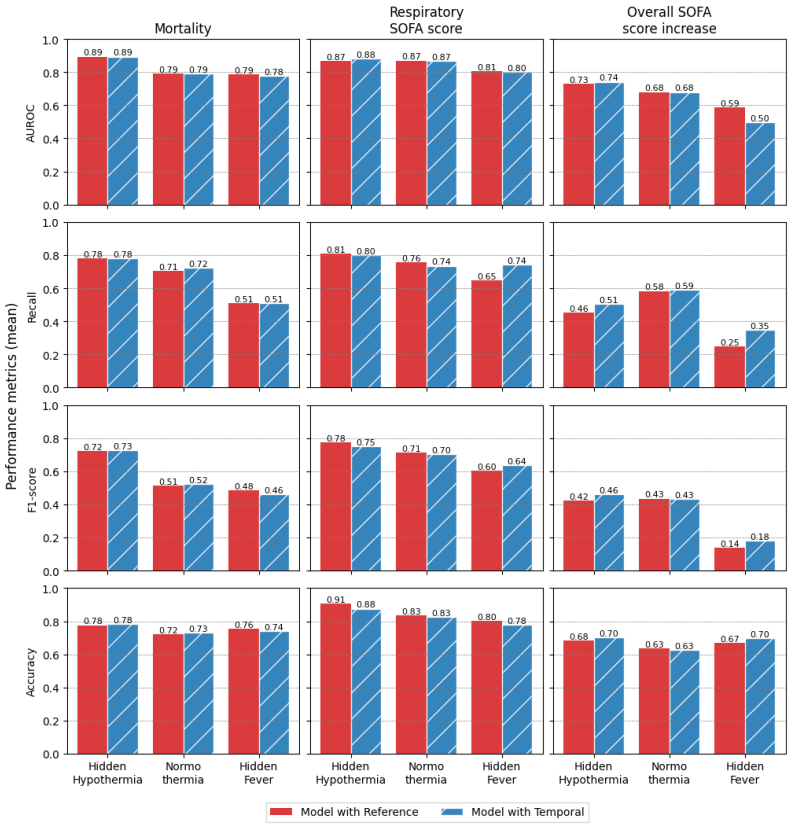


(a) LR

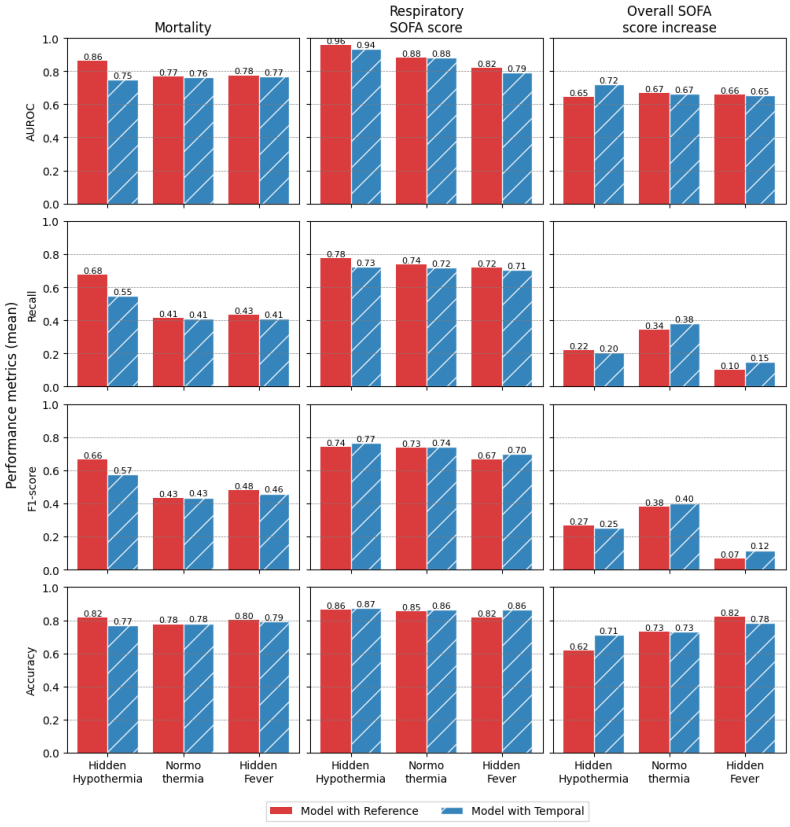


(b) XGBoost

Figure 5.12: Mean value of the XGBoost performance metrics across disparity groups, in the Thermometry study. Significant differences between reference and temporal models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001.



(a) LR



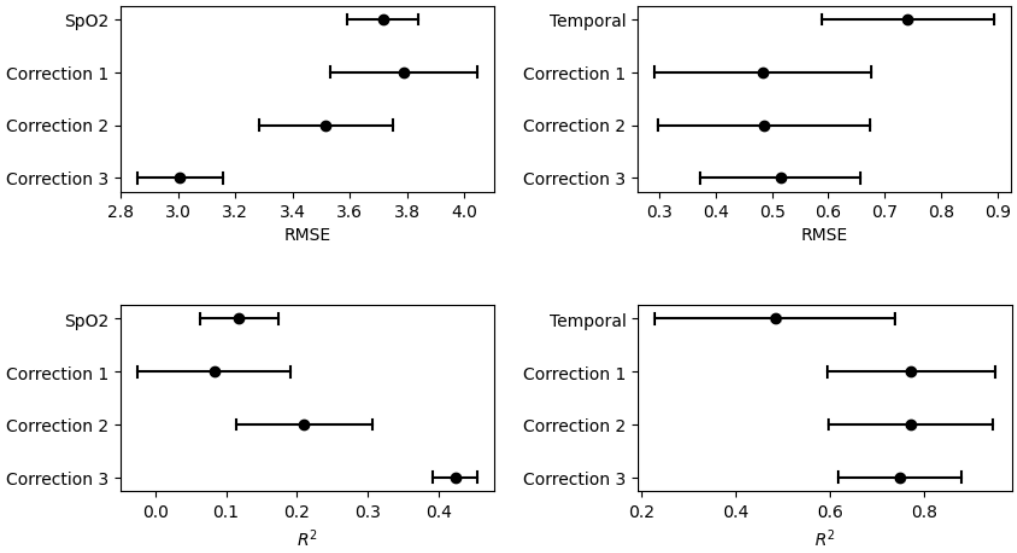
(b) XGBoost

Figure 5.13: Mean value of the performance in hidden hypothermia, normothermia and hidden fever groups. Significant differences between Temporal and Reference models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001.

5.4 Disparities Mitigation

Due to the observed disparities between SaO_2 and SpO_2 measurements and their consequent effect on the prediction model performance, three attempts of correcting the biased feature were tested. Although the temporal temperatures did not show major disparities, the rest of the methodology was still applied to the thermometry dataset, as a parallel study. Effectiveness was evaluated with RMSE and R^2. Figure 5.14 shows the metrics in the original scenario (SpO_2 and Temporal) and the changes derived from applying the correction strategies 1, 2 and 3.

The first correction model did not demonstrate an advantage in correcting SpO_2, as the RMSE increased and R^2 decreased. Model 3 exhibited better performance, and Model 2 was in between. When it comes to managing the temporal temperatures, all three strategies appeared to perform well.



(a) Blood-gas and Oximetry (b) Thermometry.

Figure 5.14: Mean and 95% CI of the RMSE and R^2 results, obtained by considering the SaO_2 and reference temperature as true values and the SpO_2 and temporal temperature (corrected or not) as predictions, in the respective studies.

5.4.1 SaO_2 vs. SpO_2

In the race and ethnicity subgroups, several situations were previously identified as significantly different. Feature correction was not advantageous because more non-consistent cases were generated with LR and XGBoost, even with Correction 3.

The disparity groups results obtained with LR were similar to the three correction approaches: most of the significant differences related to the mortality task were removed, disparities mitigation was not successful in the respiratory SOFA score and was insufficient in the last task (Figure 5.15). All models introduced new disparities. Neither of the correction models had clear advantages when used together with XGBoost.

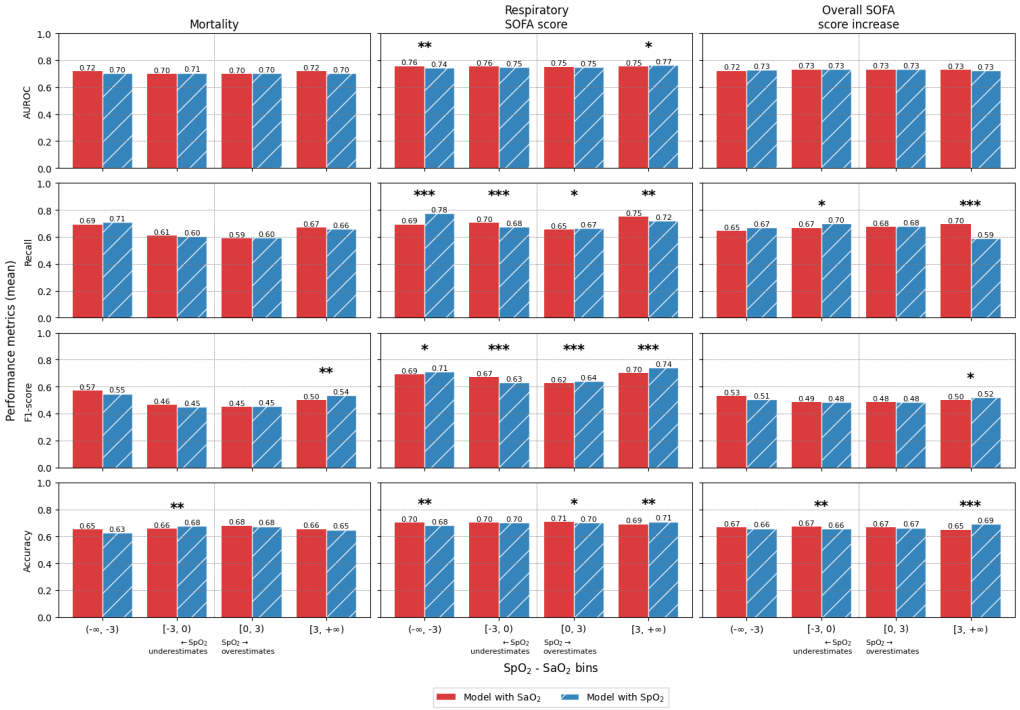


Figure 5.15: Mean values of the LR performance metrics across disparity groups, in the Blood-gas and Oximetry study with Correction 2. Significant differences between SaO_2 and SpO_2 models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001.

Unexpectedly, correction models did not spot the significant differences in the Hidden Hypoxemia cases and introduced more disparities in performance to the patients with SaO_2 and SpO_2 within the normal range.

5.4.2 Contact vs. Temporal

Results in the Thermometry dataset showed few significant differences between the Temporal and Reference datasets. However, the feature correction effect in the prediction models will be briefly presented.

Looking at the race and ethnicity subgroups, the only significant difference identified with LR was mitigated with the three strategies (Figure 5.16) but Corrections 2 and 3 introduced new differences. XGBoost results were similar, with Correction 1 also generating new differences.

In XGBoost results across disparity groups, Corrections 1 and 2 eliminated the significant differences (Figure 5.17) but Corrections 3 introduced new ones. None of the corrections was so successful when used together with LR.

In hidden hypothermia, normothermia and hidden fever groups, which had no significant disparities, LR and XGBoost performances went in the same direction: new differences were produced (Figures 5.18a and 5.18b). This unexpected situation might be caused by cases where

models failed to correct and potentially exacerbated temperature disparities. Solely Correction 3 together with XGBoost kept the original results.

Regarding the fairness metrics, there were almost no differences in Demographic parity difference and Equalized odds ratio after feature correction, for both datasets.

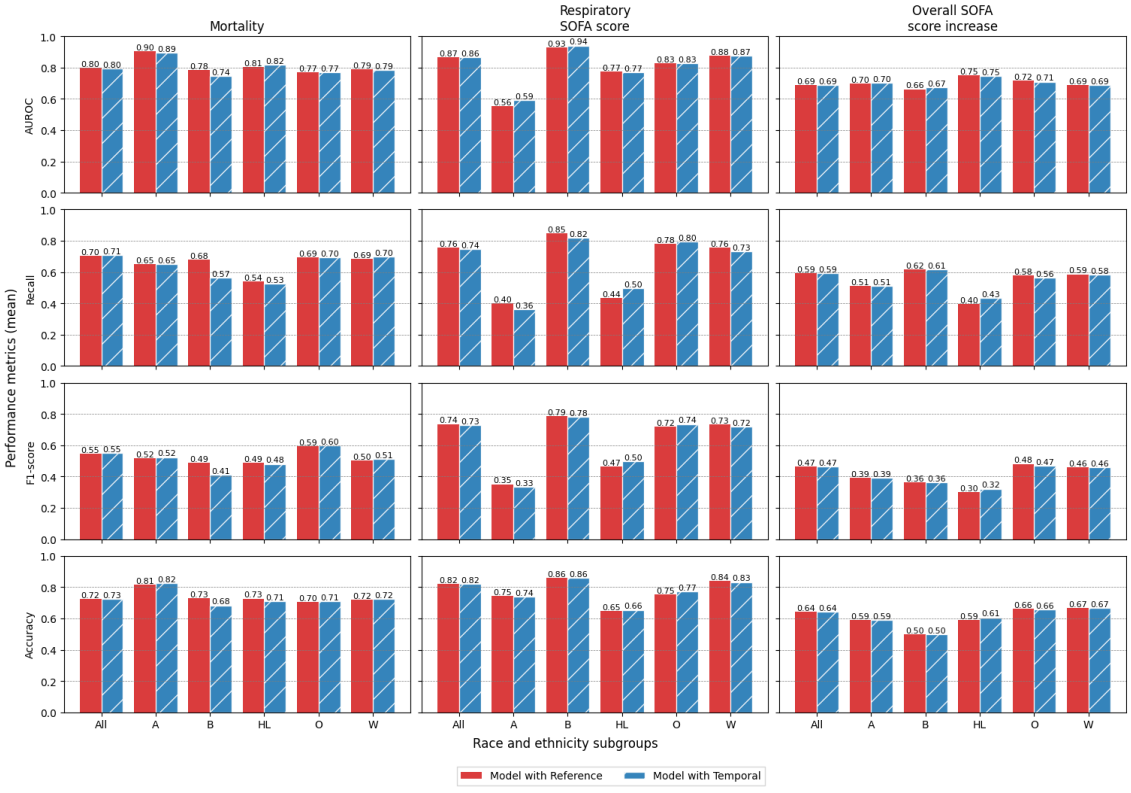


Figure 5.16: Mean values of the LR performance metrics across race and ethnicity subgroups, in the Thermometry study with Correction 1. Significant differences between Reference and Temporal models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001. A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White.

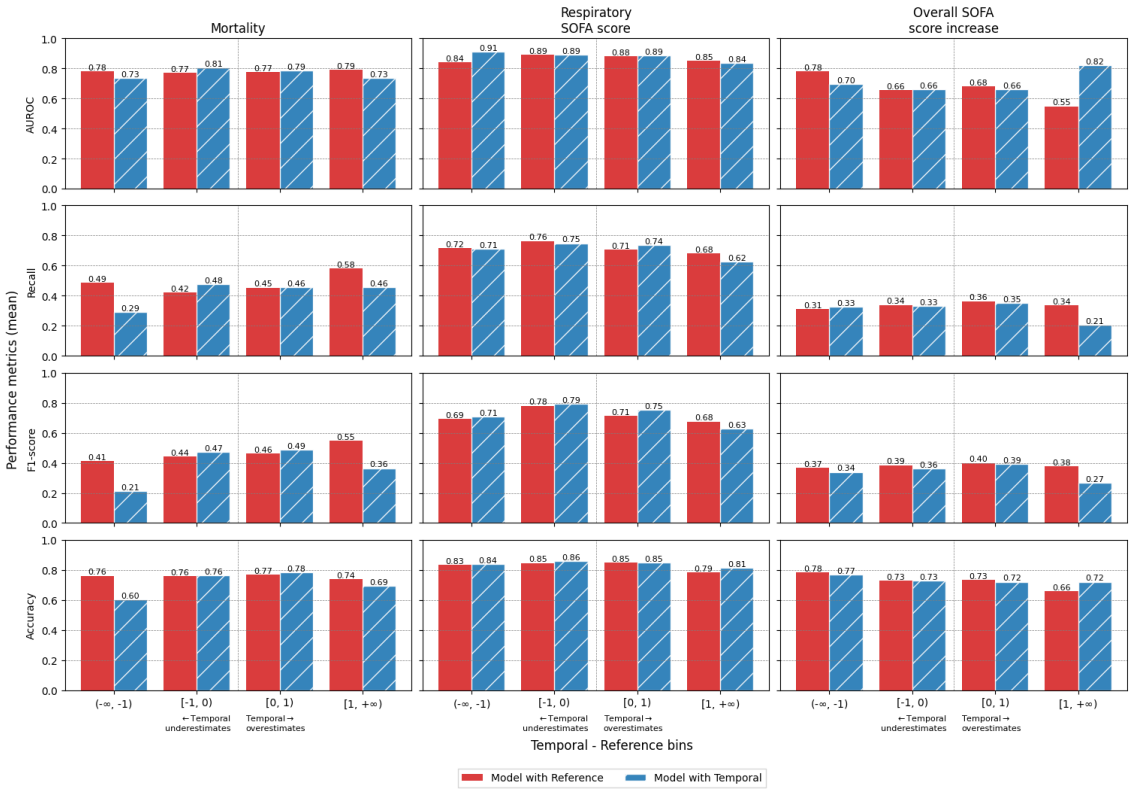
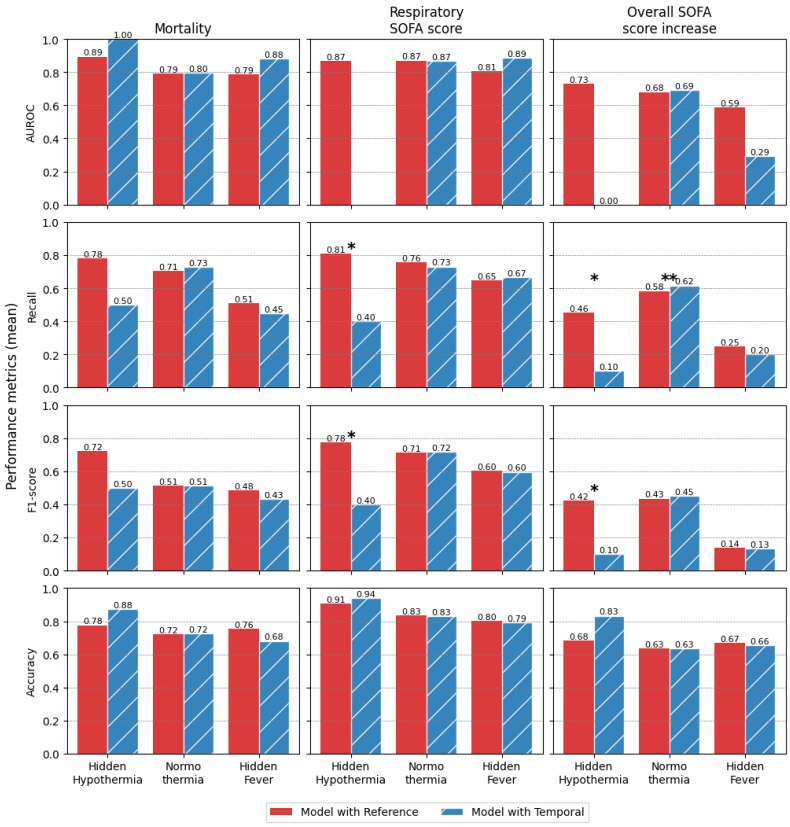
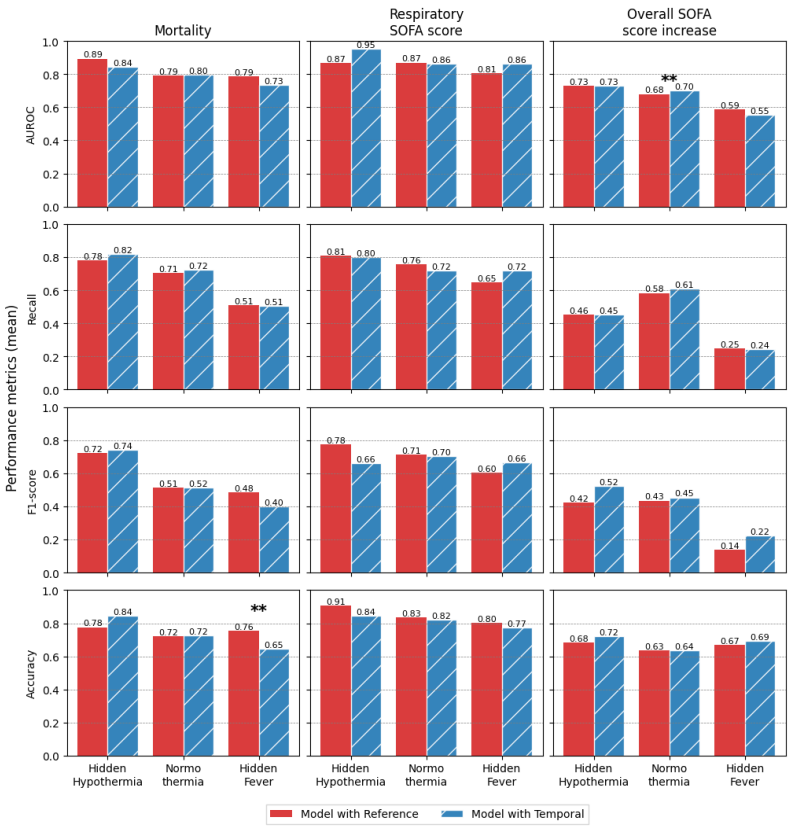


Figure 5.17: Mean values of the XGBoost performance metrics across disparity groups, in the Thermometry study with Correction 1. Significant differences between Reference and Temporal models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001.



(a) Model 1



(b) Model 3

Figure 5.18: Mean value of the LR performance in hidden hypothermia, normothermia and hidden fever groups. Significant differences between Temporal and Reference models are identified with: "$*$", for p-values ≤ 0.05; "$**$", for p-values ≤ 0.01; or "$***$", for p-values ≤ 0.001.

5.5 Summary

Both datasets had an unbalanced distribution across race and ethnicity groups, with slightly higher percentages of male subjects. The amount of available data dropped drastically when limiting the Thermometry dataset to pairs measured within a 1-hour window.

The fever prevalence results on the temperature subset showed distinct performances of temporal and oral thermometers in detecting fever, in opposite directions for Black and White patients. Among the computed OR of fever and hidden fever, solely the temporal compared to oral measurements were associated with significantly higher OR of fever in White subjects for the 37.8°C threshold.

Regarding the remaining framework, it was verified that several metrics were significantly different in multiple race and ethnicity subgroups, disparity groups and in the subgroup of patients with hidden hypoxemia, comparing the SaO_2 with the SpO_2 model. Disparities between Reference and Temporal models' performance were scarce.

The correction models seemed capable of mitigating some of the measurement bias, as RMSE and R^2 improved from the original scenario.

In the Blood-gas and Oximetry study, each of the three correction models, used together with LR, resulted in a major decrease in significant differences in the mortality task, in the disparity groups, but not completely successful. Feature correction had no advantages in the remaining approaches.

The original comparison of Temporal and Reference models has shown few significant differences. Nevertheless, three situations successfully mitigated them, without introducing new differences: Correction 1 with LR, in the race and ethnicity groups; Corrections 1 and 2 with XGBoost, in the disparity groups; and Correction 3 with XGBoost, in the hidden hypothermia, normothermia and hidden fever groups.

Chapter 6

Conclusions

Literature has shown that widely used medical devices, such as pulse oximeters, thermometers, electrocardiography machines, and sphygmomanometers, can return biased measures. This issue derives from the device's validation in populations with a lack of diversity. Input this information into machine learning models will make them learn and propagate bias. A passive mindset is an easy path but disparities will last. We must make efforts to assess and mitigate inequities in healthcare. That was the driving force of this dissertation.

The objective of curating a dataset with paired thermometry measurements was fulfilled. While the available data was demographically imbalanced with a majority of White patients, efforts are underway to expand the dataset. This might help ensure a more comprehensive interpretation of the results in different skin pigmentations. After improvements in robustness and population diversity, it is intended to make the dataset freely available. The second objective was also achieved. The designed pipeline for the assessment of disparities is easily reproduced in other datasets and for different objectives. The inherent counterfactual thinking is a useful tool to improve transparency and explainability. Although the proposed bias mitigation solutions were not as effective as anticipated, the results highlight the complexity and relevance of the topic.

6.1 Future work

Trying to find an explanation between medical device bias and skin pigmentation based on self-reported race and ethnicity is a barrier to good conclusions. Approaches with more objective and quantitative features for skin pigmentation are key.

Including more data would allow us to better validate the framework and to get more reliable conclusions. From there, the approaches for bias correction could be re-evaluated. Temperature pairs will always be scarcer due to the lower frequency of this signal acquisition, compared, for example, to oximetry. It would be interesting to build a database with bias introduced in a controlled environment to overcome this problem. This approach would allow us to better understand the effects of specific bias on the outcomes.

Afterward, it would be beneficial to use already validated models, instead of developing them from scratch, which might not be so effective. Additional studies based on a causal approach could be interesting. For example, estimating the Average Treatment Effect would help evaluate the impact of bias on outcomes.

6.2 Final remarks

We cannot be in a rush to develop technology if it can produce or augment disparities in society. Both medical devices and machine learning algorithms must pass through adequate validation to ensure equity in the results. Additionally, the audition of already deployed algorithms is crucial. However, it can only be done with transparency from entities. Obstacles in accessing data might also be a bottleneck. Literature highlights these problems but they are overlooked by most people. It is essential to bring the subject up and increase awareness, beginning within the academic environment. At the end of this journey, some steps were taken towards health equity. Besides the exploration of the topic and the development of a framework for the assessment and mitigation of medical device bias on downstream ML tasks, I had the opportunity to present the subject and methodologies at national and international conferences. Thus, another objective was fulfilled: to bring awareness to inequities in the healthcare systems and the huge impact they can have on already vulnerable populations.

References

[1] M. Schaekermann, T. Spitz, M. Pyles, H. Cole-Lewis, E. Wulczyn, S. R. Pfohl, D. Martin, R. Jaroensri, G. Keeling, Y. Liu, *et al.*, "Health equity assessment of machine learning performance (heal): a framework and dermatology ai model case study," *Eclinicalmedicine*, vol. 70, 2024.

[2] "17 goals to transform our world." https://www.un.org/sustainabledevelopment. Accessed: 2024-06-08.

[3] Y. Sun, Z. He, J. Ren, and Y. Wu, "Prediction model of in-hospital mortality in intensive care unit patients with cardiac arrest: a retrospective analysis of mimic-iv database based on machine learning," *BMC anesthesiology*, vol. 23, no. 1, p. 178, 2023.

[4] Y. Zhang, J. Hu, T. Hua, J. Zhang, Z. Zhang, and M. Yang, "Development of a machine learning-based prediction model for sepsis-associated delirium in the intensive care unit," *Scientific Reports*, vol. 13, no. 1, p. 12697, 2023.

[5] 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.

[6] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[7] "Derivative of the sigmoid function." https://towardsdatascience.com/derivative-of-the-sigmoid-function-536880cf918e. Accessed: 2024-06-11.

[8] "Decision tree." https://cio-wiki.org/wiki/Decision_Tree. Accessed: 2024-06-11.

[9] "Xgboost." https://www.geeksforgeeks.org/xgboost/. Accessed: 2024-06-11.

[10] "A gentle introduction to ml fairness metrics." https://superwise.ai/blog/gentle-introduction-ml-fairness-metrics/. Accessed: 2024-06-12.

[11] "What is pulse oximetry?." https://www.aurorahealthcare.org/services/heart-vascular/services-treatments/diagnosis-treatment-chest-lung/pulse-oximetry. Accessed: 2024-06-07.

[12] "Hand holds an infrared thermometer to measure body temperature." https://www.freepik.com/premium-vector/hand-holds-infrared-thermometer-measure-body-temperature_11882515.htm. Accessed: 2024-07-06.

[13] C. Bao, F. Deng, and S. Zhao, "Machine-learning models for prediction of sepsis patients mortality," *Medicina Intensiva (English Edition)*, vol. 47, no. 6, pp. 315–325, 2023.

[14] "Laboratory reference ranges in healthy adults." https://emedicine.medscape.com/article/2172316-overview?form=fpf. Accessed: 2024-06-12.

[15] L. Hempel, S. Sadeghi, and T. Kirsten, "Prediction of intensive care unit length of stay in the mimic-iv dataset," *Applied Sciences*, vol. 13, no. 12, p. 6930, 2023.

[16] "What is artificial intelligence (ai)?." https://www.ibm.com/topics/artificial-intelligence. Accessed: 2024-06-09.

[17] "What is ai (artificial intelligence)?." https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-ai. Accessed: 2024-06-09.

[18] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019.

[19] E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer, "An algorithmic approach to reducing unexplained pain disparities in underserved populations," *Nature Medicine*, vol. 27, no. 1, pp. 136–140, 2021.

[20] M. Rysavy, "Evidence-based medicine: a science of uncertainty and an art of probability," *AMA Journal of Ethics*, vol. 15, no. 1, pp. 4–8, 2013.

[21] A. Moran-Thomas, "How a popular medical device encodes racial bias," *Boston Review*, vol. 8, no. 5, p. 2020, 2020.

[22] M.-L. Charpignon, J. Byers, S. Cabral, L. A. Celi, C. Fernandes, J. Gallifant, M. E. Lough, D. Mlombwa, L. Moukheiber, B. A. Ong, *et al.*, "Critical bias in critical care devices," *Critical Care Clinics*, vol. 39, no. 4, pp. 795–813, 2023.

[23] A. Jubran, "Pulse oximetry," *Critical Care*, vol. 19, no. 1, 2015.

[24] M. W. Sjoding, R. P. Dickson, T. J. Iwashyna, S. E. Gay, and T. S. Valley, "Racial bias in pulse oximetry measurement," *New England Journal of Medicine*, vol. 383, no. 25, pp. 2477–2478, 2020.

[25] S. V. Bhavani, Z. Wiley, P. A. Verhoef, C. M. Coopersmith, and I. Ofotokun, "Racial differences in detection of fever using temporal vs oral temperature measurements in hospitalized patients," *Jama*, vol. 328, no. 9, pp. 885–886, 2022.

[26] P. Braveman and S. Gruskin, "Defining equity in health," *Journal of Epidemiology & Community Health*, vol. 57, no. 4, pp. 254–258, 2003.

[27] "Health equity." https://www.who.int/health-topics/health-equity#tab=tab_1. Accessed: 2024-06-08.

[28] "What is health equity?." https://www.cdc.gov/healthequity/whatis/index.html. Accessed: 2024-06-08.

[29] P. Braveman, E. Arkin, T. Orleans, D. Proctor, J. Acker, and A. Plough, "What is health equity?," *Behavioral science & policy*, vol. 4, no. 1, pp. 1–14, 2018.

[30] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[31] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eicu collaborative research database, a freely available multi-center database for critical care research," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[32] K. Alghatani, N. Ammar, A. Rezgui, A. Shaban-Nejad, *et al.*, "Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation," *JMIR medical informatics*, vol. 9, no. 5, p. e21347, 2021.

[33] Y. Chen *et al.*, "Prediction and analysis of length of stay based on nonlinear weighted xgboost algorithm in hospital," *Journal of Healthcare Engineering*, vol. 2021, 2021.

[34] R. Chen, S. Zhang, J. Li, D. Guo, W. Zhang, X. Wang, D. Tian, Z. Qu, and X. Wang, "A study on predicting the length of hospital stay for chinese patients with ischemic stroke based on the xgboost algorithm," *BMC medical informatics and decision making*, vol. 23, no. 1, p. 49, 2023.

[35] A. Kakadiaris, "Evaluating the fairness of the mimic-iv dataset and a baseline algorithm: Application to the icu length of stay prediction," *arXiv preprint arXiv:2401.00902*, 2023.

[36] L. F. Nakayama, J. Matos, J. Quion, F. Novaes, W. G. Mitchell, R. Mwavu, J.-Y. J. Hung, W. Phanphruk, J. S. Cardoso, L. A. Celi, *et al.*, "Unmasking biases and navigating pitfalls in the ophthalmic artificial intelligence lifecycle: A review," *arXiv preprint arXiv:2310.04997*, 2023.

[37] L. H. Nazer, R. Zatarah, S. Waldrip, J. X. C. Ke, M. Moukheiber, A. K. Khanna, R. S. Hicklen, L. Moukheiber, D. Moukheiber, H. Ma, *et al.*, "Bias in artificial intelligence algorithms and recommendations for mitigation," *PLOS digital health*, vol. 2, no. 6, p. e0000278, 2023.

[38] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[39] H. Ledford, "Millions of black people affected by racial bias in health-care algorithms. 2019."

[40] A. Balagopalan, I. Baldini, L. A. Celi, J. Gichoya, L. G. McCoy, T. Naumann, U. Shalit, M. van der Schaar, and K. L. Wagstaff, "Machine learning for healthcare that matters: Reorienting from technical novelty to equitable impact," *PLOS Digital Health*, vol. 3, no. 4, p. e0000474, 2024.

[41] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[42] R. Blankstein, R. P. Ward, M. Arnsdorf, B. Jones, Y.-B. Lou, and M. Pine, "Female gender is an independent predictor of operative mortality after coronary artery bypass graft surgery: contemporary analysis of 31 midwestern hospitals," *Circulation*, vol. 112, no. 9_supplement, pp. I–323, 2005.

[43] S. Boekholdt, F. Sacks, J. Jukema, J. Shepherd, D. Freeman, A. McMahon, F. Cambien, V. Nicaud, G. De Grooth, P. Talmud, *et al.*, "Cholesteryl ester transfer protein taqib variant, high-density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of pravastatin treatment: individual patient meta-analysis of 13 677 subjects," *Circulation*, vol. 111, no. 3, pp. 278–287, 2005.

[44] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[45] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[46] 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*, pp. 785–794, 2016.

[47] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[48] M. Gupta, B. Gallamoza, N. Cutrona, P. Dhakal, R. Poulain, and R. Beheshti, "An extensive data processing pipeline for mimic-iv," in *Machine Learning for Health*, pp. 311–325, PMLR, 2022.

[49] "Regression metrics." https://permetrics.readthedocs.io/en/latest/pages/regression.html. Accessed: 2024-06-12.

[50] "Counterfactuals in machine learning: Exploring the power of "what if?"." https://medium.com/aimonks/counterfactuals-in-machine-learning-exploring-the-power-of-what-if-b210934648e. Accessed: 2024-06-11.

[51] S.-I. Lee and E. J. Topol, "The clinical potential of counterfactual ai models," *The Lancet*, vol. 403, no. 10428, p. 717, 2024.

[52] D. Castro, S. Patil, M. Zubair, and M. Keenaghan, "Arterial blood gas," *StatPearls*, 2024.

[53] "Hypoxia what is hypoxia?." https://my.clevelandclinic.org/health/diseases/23063-hypoxia. Accessed: 2024-06-23.

[54] P. Sirohiya, A. Elavarasi, H. K. R. Sagiraju, M. Baruah, N. Gupta, R. K. Garg, S. S. Paul, B. K. Ratre, R. Singh, B. Kumar, *et al.*, "Silent hypoxia in coronavirus disease-2019: Is it more dangerous?-a retrospective cohort study," *Lung India*, vol. 39, no. 3, pp. 247–253, 2022.

[55] "Medical electrical equipment particular requirements for basic safety and essential performance of pulse oximeter equipment." https://www.iso.org/obp/ui/#iso:std:iso:80601:-2-61:ed-2:v2:en. Accessed: 2024-06-07.

[56] K. Poorzargar, C. Pham, J. Ariaratnam, K. Lee, M. Parotto, M. Englesakis, F. Chung, and M. Nagappa, "Accuracy of pulse oximeters in measuring oxygen saturation in patients with poor peripheral perfusion: a systematic review," *Journal of clinical monitoring and computing*, vol. 36, no. 4, pp. 961–973, 2022.

[57] A.-K. I. Wong, M. Charpignon, H. Kim, C. Josef, A. A. De Hond, J. J. Fojas, A. Tabaie, X. Liu, E. Mireles-Cabodevila, L. Carvalho, *et al.*, "Analysis of discrepancies between pulse

oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality," *JAMA Network Open*, vol. 4, no. 11, pp. e2131674–e2131674, 2021.

[58] P. E. Bickler, J. R. Feiner, and J. W. Severinghaus, "Effects of skin pigmentation on pulse oximeter accuracy at low saturation," *The Journal of the American Society of Anesthesiologists*, vol. 102, no. 4, pp. 715–719, 2005.

[59] J. R. Feiner, J. W. Severinghaus, and P. E. Bickler, "Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender," *Anesthesia & Analgesia*, vol. 105, no. 6, pp. S18–S23, 2007.

[60] S. K. N. Swamy, C. He, B. R. Hayes-Gill, D. J. Clark, S. Green, and S. P. Morgan, "Pulse oximeter bench tests under different simulated skin tones," *Medical & Biological Engineering & Computing*, pp. 1–13, 2024.

[61] B. Chacko and J. Peter, "Temperature monitoring in the intensive care unit," *Indian Journal of Respiratory Care*, vol. 7, p. 28, 01 2018.

[62] S. Tharakan, K. Nomoto, S. Miyashita, and K. Ishikawa, "Body temperature correlates with mortality in covid-19 patients," *Critical care*, vol. 24, pp. 1–3, 2020.

[63] B. Circiumaru, G. Baldock, and J. Cohen, "A prospective study of fever in the intensive care unit," *Intensive Care Medicine*, vol. 25, pp. 668–673, 1999.

[64] S. Dugar, C. Choudhary, and A. Duggal, "Sepsis and septic shock: Guideline-based management," *Cleve Clin J Med*, vol. 87, no. 1, pp. 53–64, 2020.

[65] M. Doman, M. Thy, J. Dessajan, M. Dlela, H. Do Rego, E. Cariou, M. Ejzenberg, L. Bouadma, E. de Montmollin, and J.-F. Timsit, "Temperature control in sepsis," *Frontiers in Medicine*, vol. 10, p. 1292468, 2023.

[66] 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.

[67] D. O. Thomas-Rüddel, P. Hoffmann, D. Schwarzkopf, C. Scheer, F. Bach, M. Komann, H. Gerlach, M. Weiss, M. Lindner, H. Rüddel, *et al.*, "Fever and hypothermia represent two populations of sepsis patients and are associated with outside temperature," *Critical Care*, vol. 25, pp. 1–10, 2021.

[68] S. V. Bhavani, K. A. Carey, E. R. Gilbert, M. Afshar, P. A. Verhoef, and M. M. Churpek, "Identifying novel sepsis subphenotypes using temperature trajectories," *American journal of respiratory and critical care medicine*, vol. 200, no. 3, pp. 327–335, 2019.

[69] D. Peres Bota, F. Lopes Ferreira, C. Mélot, and J. L. Vincent, "Body temperature alterations in the critically ill," *Intensive care medicine*, vol. 30, pp. 811–816, 2004.

[70] K. B. Laupland, J.-R. Zahar, C. Adrie, C. Schwebel, D. Goldgran-Toledano, E. Azoulay, M. Garrouste-Orgeas, Y. Cohen, S. Jamali, B. Souweine, *et al.*, "Determinants of temperature abnormalities and influence on outcome of critical illness," *Critical care medicine*, vol. 40, no. 1, pp. 145–151, 2012.

[71] Z. Rumbus, R. Matics, P. Hegyi, C. Zsiboras, I. Szabo, A. Illes, E. Petervari, M. Balasko, K. Marta, A. Miko, *et al.*, "Fever is associated with reduced, hypothermia with increased mortality in septic patients: a meta-analysis of clinical trials," *PloS one*, vol. 12, no. 1, p. e0170152, 2017.

[72] R. Sanga, S. Zanotti, C. Schorr, B. Milcareck, K. Hunter, P. Dellinger, and J. Parrilo, "Relation between temperature in the initial 24 hours in patients with severe sepsis or septic shock with mortality and length of stay in the icu," *Critical Care*, vol. 16, pp. 1–189, 2012.

[73] M. Saxena, P. Young, D. Pilcher, M. Bailey, D. Harrison, R. Bellomo, S. Finfer, R. Beasley, J. Hyam, D. Menon, *et al.*, "Early temperature and mortality in critically ill patients with acute neurological diseases: trauma and stroke differ from infection," *Intensive care medicine*, vol. 41, pp. 823–832, 2015.

[74] D. J. Tan, J. Chen, Y. Zhou, J. S. Q. Ong, R. J. X. Sin, T. V. Bui, A. A. Mehta, M. Feng, and K. C. See, "Association of body temperature and mortality in critically ill patients: an observational study using two large databases," *European Journal of Medical Research*, vol. 29, no. 1, p. 33, 2024.

[75] S. L. Cutuli, E. J. See, E. A. Osawa, P. Ancona, D. Marshall, G. M. Eastwood, N. J. Glassford, and R. Bellomo, "Accuracy of non-invasive body temperature measurement methods in adult patients admitted to the intensive care unit: a systematic review and meta-analysis," *Critical Care and Resuscitation*, vol. 23, no. 1, pp. 6–13, 2021.

[76] B. Chacko, J. Peter, *et al.*, "Temperature monitoring in the intensive care unit," *Indian J Respir Care*, vol. 7, no. 1, p. 28, 2018.

[77] F. Pompei and M. Pompei, "Noninvasive temporal artery thermometry: physics, physiology, and clinical accuracy," in *Thermosense XXVI*, vol. 5405, pp. 61–67, SPIE, 2004.

[78] C. A. Dinarello, "Thermoregulation and the pathogenesis of fever," *Infectious Disease Clinics*, vol. 10, no. 2, pp. 433–449, 1996.

[79] T. Holder, F. S. W. Hooper, D. Yates, Z. Tse, S. Patil, A. Moussa, L. Batten, V. Radhakrishnan, M. Allison, C. Hewitt, *et al.*, "Clinical accuracy of infrared temperature measurement devices: a comparison against non-invasive core-body temperature," *Clinical Medicine*, vol. 23, no. 2, pp. 157–163, 2023.

[80] M. Charlton, S. A. Stanley, Z. Whitman, V. Wenn, T. J. Coats, M. Sims, and J. P. Thompson, "The effect of constitutive pigmentation on the measured emissivity of human skin," *Plos one*, vol. 15, no. 11, p. e0241843, 2020.

[81] S. Farnell, L. Maxwell, S. Tan, A. Rhodes, and B. Philips, "Temperature measurement: Comparison of non-invasive methods used in adult critical care," *Journal of clinical nursing*, vol. 14, no. 5, pp. 632–639, 2005.

[82] K. Dempsey, M. Lindsay, J. E. Tcheng, and A.-K. I. Wong, "The high price of equity in pulse oximetry: A cost evaluation and need for interim solutions," *medRxiv*, 2023.

[83] J. Matos, T. Struja, J. Gallifant, M.-L. Charpignon, J. S. Cardoso, and L. A. Celi, "Shining light on dark skin: Pulse oximetry correction models," in *2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG)*, pp. 211–214, IEEE, 2023.

[84] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[85] J. Matos, T. Struja, J. Gallifant, L. Nakayama, M.-L. Charpignon, X. Liu, N. Economou-Zavlanos, J. S. Cardoso, K. S. Johnson, N. Bhavsar, *et al.*, "Bold: Blood-gas and oximetry linked dataset," *Scientific Data*, vol. 11, no. 1, p. 535, 2024.

[86] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[87] C. Van Walraven, P. C. Austin, A. Jennings, H. Quan, and A. J. Forster, "A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data," *Medical care*, vol. 47, no. 6, pp. 626–633, 2009.

[88] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation," *Journal of chronic diseases*, vol. 40, no. 5, pp. 373–383, 1987.

[89] M. Szumilas, "Explaining odds ratios," *Journal of the Canadian academy of child and adolescent psychiatry*, vol. 19, no. 3, p. 227, 2010.

Appendix A

Datasets information

A.1 Flow Diagrams

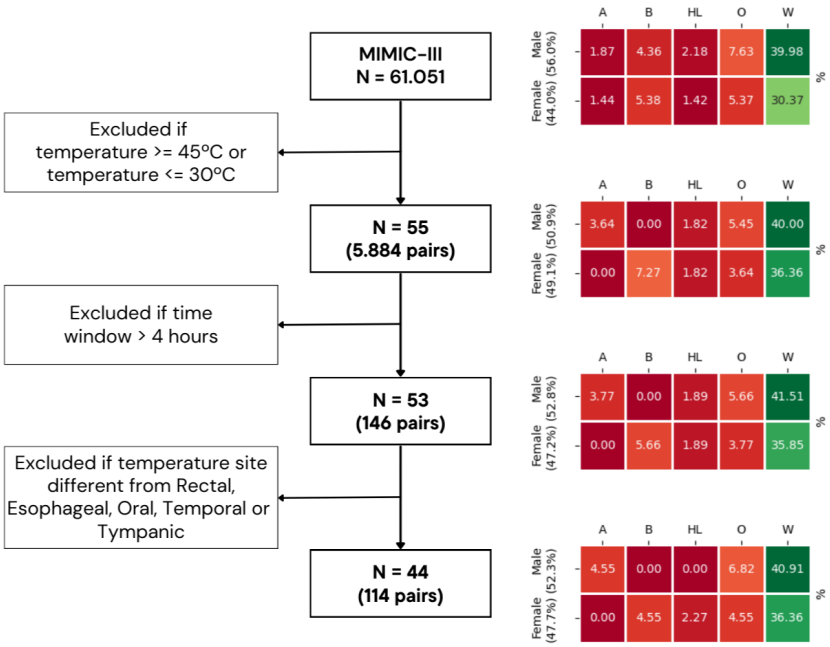


Figure A.1: MIMIC-III flow diagram. The first box displays the number of subjects (N) in the original database and the subsequent ones show the available pairs considering exclusion criteria. The race/ethnicity groups are indicated by their initials (A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White).

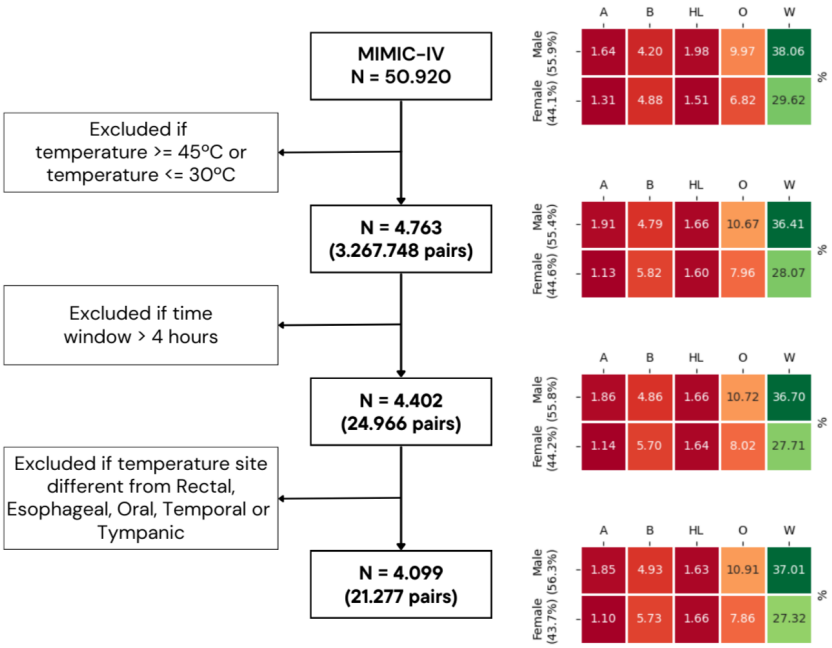


Figure A.2: MIMIC-IV flow diagram. The first box displays the number of subjects (N) in the original database and the subsequent ones show the available pairs considering exclusion criteria. The race/ethnicity groups are indicated by their initials (A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White).

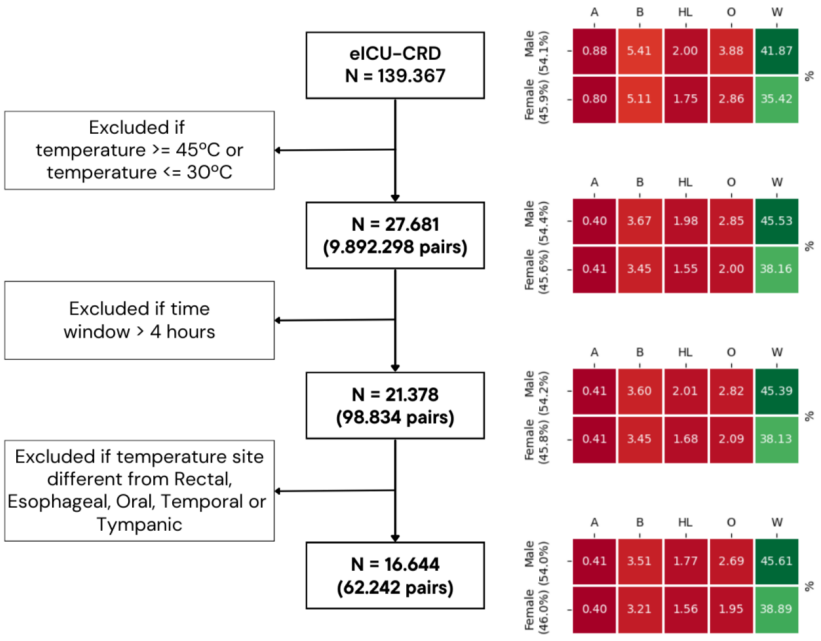


Figure A.3: eICU-CRD flow diagram. The first box displays the number of subjects (N) in the original database and the subsequent ones show the available pairs considering exclusion criteria. The race/ethnicity groups are indicated by their initials (A: Asian; B: Black; HL: Hispanic or Latino; O: Other; W: White).

Appendix B

Complete results

In the following tables, race and ethnicity groups are represented as: A - Asian; B - Black; HL - Hispanic or Latino; O - Other; W - White. The five scenarios are represented as: R - Reference; T - Temporal; C1 - Correction 1; C2 - Correction 2; C3 - Correction 3.

B.1 Blood-gas and Oximetry

B.1.1 Training set

Table B.1: Performance metrics in the training set per race and ethnicity groups, in the Blood-gas and Oximetry study.

Logistic Regression

Task	Feature	AUROC						Recall						F1-score						Accuracy					
		All	A	B	HL	O	W	All	A	B	HL	O	W	All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	Reference	0.711	0.719	0.705	0.748	0.736	0.706	0.625	0.606	0.591	0.666	0.672	0.623	0.476	0.53	0.471	0.554	0.484	0.468	0.67	0.685	0.679	0.692	0.676	0.666
	SpO2	0.71	0.716	0.704	0.748	0.734	0.706	0.627	0.594	0.586	0.672	0.673	0.626	0.475	0.524	0.469	0.557	0.483	0.468	0.668	0.683	0.679	0.693	0.674	0.663
	Correction 1	0.71	0.723	0.704	0.747	0.737	0.706	0.625	0.612	0.592	0.669	0.673	0.622	0.476	0.536	0.47	0.554	0.485	0.468	0.669	0.689	0.677	0.691	0.677	0.666
	Correction 2	0.71	0.724	0.703	0.748	0.736	0.706	0.625	0.609	0.593	0.671	0.672	0.622	0.476	0.536	0.469	0.558	0.485	0.468	0.67	0.69	0.676	0.694	0.676	0.666
	Correction 3	0.712	0.727	0.706	0.747	0.738	0.707	0.623	0.619	0.587	0.661	0.673	0.621	0.477	0.542	0.471	0.554	0.487	0.469	0.672	0.693	0.682	0.694	0.679	0.668
Respiratory SOFA score	Reference	0.759	0.801	0.779	0.768	0.768	0.753	0.693	0.724	0.706	0.706	0.693	0.689	0.659	0.687	0.661	0.681	0.666	0.656	0.702	0.741	0.721	0.707	0.711	0.697
	SpO2	0.755	0.796	0.772	0.766	0.767	0.75	0.688	0.719	0.697	0.702	0.688	0.685	0.657	0.683	0.659	0.68	0.664	0.653	0.701	0.738	0.722	0.707	0.711	0.695
	Correction 1	0.754	0.796	0.776	0.764	0.764	0.748	0.689	0.723	0.707	0.706	0.687	0.685	0.657	0.684	0.661	0.68	0.662	0.653	0.7	0.738	0.72	0.705	0.708	0.695
	Correction 2	0.755	0.796	0.777	0.765	0.765	0.749	0.69	0.722	0.706	0.707	0.689	0.686	0.657	0.683	0.661	0.681	0.664	0.654	0.701	0.737	0.721	0.707	0.709	0.696
	Correction 3	0.755	0.795	0.777	0.765	0.764	0.749	0.688	0.723	0.702	0.7	0.688	0.685	0.657	0.685	0.66	0.679	0.665	0.654	0.702	0.74	0.721	0.707	0.711	0.696
Overall SOFA score increase	Reference	0.732	0.704	0.725	0.724	0.734	0.734	0.679	0.679	0.679	0.662	0.676	0.681	0.491	0.448	0.485	0.482	0.504	0.493	0.665	0.619	0.66	0.671	0.669	0.666
	SpO2	0.731	0.702	0.723	0.722	0.734	0.734	0.679	0.671	0.67	0.657	0.674	0.682	0.491	0.444	0.483	0.478	0.503	0.493	0.665	0.617	0.662	0.668	0.668	0.665
	Correction 1	0.731	0.703	0.724	0.72	0.734	0.733	0.678	0.676	0.68	0.661	0.673	0.68	0.49	0.442	0.483	0.478	0.501	0.492	0.664	0.612	0.657	0.666	0.667	0.666
	Correction 2	0.731	0.703	0.724	0.721	0.733	0.733	0.678	0.676	0.68	0.66	0.675	0.68	0.49	0.444	0.483	0.477	0.503	0.492	0.664	0.615	0.658	0.666	0.667	0.666
	Correction 3	0.731	0.699	0.723	0.721	0.733	0.733	0.678	0.672	0.676	0.66	0.673	0.68	0.49	0.439	0.482	0.478	0.501	0.492	0.664	0.609	0.657	0.666	0.667	0.665

Random Forest

Task	Feature	AUROC						Recall						F1-score						Accuracy					
		All	A	B	HL	O	W	All	A	B	HL	O	W	All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

XGBoost

Task	Feature	AUROC						Recall						F1-score						Accuracy					
		All	A	B	HL	O	W	All	A	B	HL	O	W	All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	Reference	0.887	0.905	0.894	0.895	0.893	0.884	0.783	0.821	0.773	0.826	0.800	0.778	0.661	0.725	0.681	0.703	0.656	0.653	0.807	0.817	0.825	0.800	0.810	0.805
	SpO2	0.887	0.901	0.895	0.896	0.891	0.884	0.785	0.813	0.778	0.824	0.794	0.781	0.662	0.715	0.684	0.703	0.650	0.655	0.807	0.810	0.826	0.800	0.806	0.805
	Correction 1	0.889	0.905	0.896	0.898	0.895	0.886	0.787	0.815	0.780	0.823	0.800	0.782	0.665	0.724	0.686	0.704	0.659	0.658	0.810	0.818	0.827	0.802	0.813	0.808
	Correction 2	0.889	0.908	0.897	0.898	0.895	0.887	0.788	0.822	0.781	0.825	0.802	0.783	0.665	0.726	0.686	0.705	0.658	0.658	0.810	0.818	0.827	0.802	0.811	0.808
	Correction 3	0.887	0.904	0.894	0.895	0.893	0.885	0.784	0.820	0.778	0.823	0.797	0.780	0.663	0.723	0.684	0.704	0.657	0.655	0.809	0.815	0.826	0.801	0.811	0.806
Respiratory SOFA score	Reference	0.847	0.877	0.868	0.855	0.859	0.842	0.735	0.768	0.755	0.722	0.756	0.730	0.726	0.751	0.739	0.738	0.742	0.722	0.770	0.800	0.794	0.774	0.781	0.764
	SpO2	0.845	0.875	0.865	0.854	0.857	0.839	0.733	0.757	0.739	0.726	0.754	0.730	0.724	0.747	0.733	0.741	0.737	0.719	0.767	0.799	0.792	0.776	0.777	0.761
	Correction 1	0.846	0.875	0.868	0.853	0.859	0.841	0.737	0.767	0.761	0.727	0.756	0.732	0.725	0.750	0.736	0.738	0.740	0.721	0.768	0.800	0.790	0.772	0.779	0.762
	Correction 2	0.847	0.874	0.869	0.854	0.859	0.841	0.737	0.767	0.762	0.726	0.757	0.732	0.725	0.750	0.737	0.739	0.740	0.721	0.768	0.800	0.790	0.773	0.779	0.762
	Correction 3	0.846	0.876	0.869	0.854	0.859	0.840	0.737	0.769	0.761	0.723	0.758	0.731	0.724	0.751	0.737	0.737	0.740	0.720	0.767	0.800	0.790	0.772	0.779	0.761
Overall SOFA score increase	Reference	0.857	0.842	0.858	0.865	0.862	0.857	0.767	0.730	0.791	0.769	0.777	0.763	0.618	0.586	0.610	0.623	0.633	0.618	0.774	0.765	0.762	0.785	0.776	0.775
	SpO2	0.856	0.840	0.857	0.864	0.862	0.856	0.766	0.721	0.784	0.765	0.777	0.764	0.616	0.584	0.612	0.622	0.632	0.615	0.773	0.766	0.766	0.786	0.775	0.773
	Correction 1	0.858	0.846	0.859	0.864	0.862	0.857	0.768	0.741	0.797	0.767	0.777	0.764	0.618	0.594	0.612	0.619	0.635	0.618	0.774	0.770	0.761	0.782	0.778	0.775
	Correction 2	0.858	0.844	0.860	0.865	0.864	0.857	0.768	0.743	0.798	0.768	0.776	0.764	0.619	0.594	0.612	0.621	0.635	0.618	0.775	0.769	0.762	0.784	0.778	0.776
	Correction 3	0.858	0.844	0.860	0.864	0.862	0.857	0.767	0.738	0.797	0.768	0.773	0.763	0.618	0.590	0.612	0.621	0.632	0.617	0.774	0.767	0.762	0.783	0.776	0.775

Table B.2: Performance metrics in the training set per disparity groups, in the Blood-gas and Oximetry study.

Logistic Regression

Task	Feature	AUROC				Recall				F1-score				Accuracy			
		(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)
In-hospital mortality	Reference	0.725	0.703	0.703	0.720	0.694	0.615	0.594	0.672	0.576	0.469	0.451	0.505	0.654	0.661	0.680	0.657
	SpO2	0.726	0.703	0.703	0.718	0.760	0.641	0.597	0.643	0.579	0.470	0.450	0.502	0.626	0.648	0.678	0.667
	Correction 1	0.709	0.701	0.710	0.703	0.706	0.624	0.598	0.646	0.553	0.472	0.444	0.525	0.634	0.657	0.684	0.653
	Correction 2	0.708	0.709	0.705	0.707	0.714	0.606	0.600	0.664	0.548	0.453	0.456	0.540	0.628	0.677	0.677	0.652
	Correction 3	0.706	0.706	0.716	0.712	0.680	0.604	0.624	0.673	0.526	0.458	0.475	0.544	0.639	0.674	0.677	0.656
Respiratory SOFA score	Reference	0.758	0.756	0.753	0.755	0.694	0.704	0.654	0.749	0.692	0.670	0.625	0.702	0.702	0.704	0.707	0.690
	SpO2	0.752	0.756	0.752	0.757	0.781	0.726	0.657	0.699	0.700	0.672	0.624	0.694	0.677	0.697	0.705	0.698
	Correction 1	0.743	0.753	0.749	0.764	0.774	0.709	0.656	0.695	0.714	0.668	0.620	0.714	0.682	0.697	0.704	0.703
	Correction 2	0.743	0.753	0.751	0.765	0.774	0.676	0.666	0.721	0.714	0.631	0.640	0.740	0.682	0.703	0.702	0.708
	Correction 3	0.727	0.756	0.751	0.750	0.750	0.681	0.677	0.719	0.680	0.628	0.669	0.734	0.668	0.707	0.701	0.698
Overall SOFA score increase	Reference	0.723	0.733	0.733	0.731	0.644	0.668	0.677	0.698	0.530	0.487	0.484	0.502	0.666	0.674	0.669	0.651
	SpO2	0.725	0.732	0.733	0.729	0.736	0.698	0.681	0.649	0.534	0.484	0.484	0.500	0.625	0.656	0.667	0.672
	Correction 1	0.731	0.730	0.736	0.727	0.676	0.659	0.710	0.589	0.509	0.480	0.486	0.513	0.661	0.676	0.651	0.696
	Correction 2	0.730	0.735	0.732	0.726	0.677	0.700	0.679	0.593	0.507	0.487	0.482	0.521	0.659	0.660	0.664	0.692
	Correction 3	0.734	0.737	0.727	0.721	0.667	0.689	0.679	0.628	0.510	0.478	0.490	0.544	0.671	0.666	0.658	0.671

Random Forest

Task	Feature	AUROC				Recall				F1-score				Accuracy			
		(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)
In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

XGBoost

Task	Feature	AUROC				Recall				F1-score				Accuracy			
		(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)
In-hospital mortality	Reference	0.907	0.882	0.879	0.899	0.853	0.779	0.755	0.823	0.747	0.657	0.637	0.688	0.805	0.801	0.810	0.806
	SpO2	0.909	0.882	0.879	0.899	0.898	0.788	0.758	0.808	0.732	0.653	0.638	0.696	0.777	0.796	0.810	0.816
	Correction 1	0.905	0.883	0.881	0.907	0.875	0.781	0.751	0.837	0.726	0.662	0.632	0.727	0.788	0.804	0.815	0.813
	Correction 2	0.904	0.883	0.884	0.906	0.868	0.763	0.771	0.832	0.723	0.642	0.649	0.732	0.790	0.812	0.812	0.812
	Correction 3	0.886	0.883	0.891	0.893	0.836	0.769	0.784	0.825	0.693	0.648	0.663	0.715	0.782	0.810	0.814	0.800
Respiratory SOFA score	Reference	0.861	0.841	0.846	0.847	0.735	0.711	0.692	0.820	0.760	0.721	0.698	0.767	0.777	0.765	0.776	0.756
	SpO2	0.867	0.840	0.845	0.845	0.864	0.779	0.698	0.739	0.775	0.728	0.698	0.754	0.758	0.751	0.774	0.764
	Correction 1	0.845	0.842	0.843	0.857	0.848	0.751	0.696	0.759	0.774	0.727	0.694	0.781	0.746	0.758	0.775	0.772
	Correction 2	0.844	0.845	0.845	0.852	0.842	0.725	0.706	0.763	0.772	0.704	0.710	0.790	0.745	0.771	0.771	0.768
	Correction 3	0.825	0.850	0.841	0.839	0.823	0.719	0.730	0.780	0.735	0.700	0.733	0.792	0.720	0.777	0.762	0.762
Overall SOFA score increase	Reference	0.867	0.853	0.857	0.859	0.778	0.730	0.757	0.809	0.674	0.608	0.612	0.624	0.780	0.782	0.780	0.754
	SpO2	0.869	0.852	0.857	0.858	0.879	0.778	0.757	0.755	0.642	0.599	0.611	0.634	0.713	0.760	0.779	0.780
	Correction 1	0.870	0.857	0.853	0.872	0.836	0.742	0.770	0.757	0.637	0.612	0.605	0.670	0.752	0.787	0.767	0.797
	Correction 2	0.868	0.856	0.854	0.872	0.830	0.768	0.756	0.748	0.637	0.612	0.607	0.677	0.755	0.775	0.774	0.798
	Correction 3	0.862	0.861	0.851	0.859	0.813	0.764	0.761	0.778	0.628	0.608	0.613	0.679	0.753	0.782	0.767	0.771

Table B.3: Performance metrics in the training set per No/Hidden Hypoxemia groups, in the Blood-gas and Oximetry study.

Logistic Regression

Task	Feature	AUROC		Recall		F1-score		Accuracy	
		No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia
In-hospital mortality	Reference	0.707	0.728	0.611	0.781	0.464	0.584	0.672	0.628
	SpO2	0.706	0.733	0.614	0.730	0.464	0.586	0.669	0.655
	Correction 1	0.705	0.714	0.609	0.711	0.460	0.581	0.672	0.647
	Correction 2	0.705	0.712	0.606	0.714	0.460	0.580	0.673	0.644
	Correction 3	0.709	0.711	0.613	0.702	0.468	0.569	0.673	0.647
Respiratory SOFA score	Reference	0.754	0.730	0.680	0.853	0.649	0.754	0.703	0.661
	SpO2	0.751	0.767	0.679	0.728	0.648	0.753	0.701	0.709
	Correction 1	0.750	0.756	0.677	0.718	0.645	0.758	0.701	0.704
	Correction 2	0.750	0.757	0.676	0.714	0.645	0.755	0.702	0.703
	Correction 3	0.751	0.743	0.679	0.735	0.648	0.753	0.702	0.701
Overall SOFA score increase	Reference	0.732	0.709	0.673	0.749	0.484	0.577	0.668	0.613
	SpO2	0.731	0.716	0.677	0.653	0.483	0.571	0.666	0.654
	Correction 1	0.731	0.720	0.676	0.617	0.482	0.572	0.665	0.669
	Correction 2	0.732	0.720	0.676	0.622	0.482	0.575	0.666	0.670
	Correction 3	0.732	0.711	0.678	0.636	0.484	0.565	0.665	0.656

Random Forest

Task	Feature	AUROC		Recall		F1-score		Accuracy	
		No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia
In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SpO2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

XGBoost

Task	Feature	AUROC		Recall		F1-score		Accuracy	
		No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia
In-hospital mortality	Reference	0.884	0.921	0.772	0.903	0.652	0.742	0.808	0.790
	SpO2	0.884	0.911	0.775	0.868	0.653	0.745	0.808	0.801
	Correction 1	0.885	0.906	0.772	0.859	0.653	0.750	0.811	0.802
	Correction 2	0.885	0.905	0.773	0.853	0.653	0.746	0.811	0.800
	Correction 3	0.885	0.897	0.777	0.856	0.656	0.724	0.809	0.784
Respiratory SOFA score	Reference	0.843	0.878	0.720	0.919	0.715	0.839	0.768	0.786
	SpO2	0.841	0.863	0.722	0.828	0.713	0.823	0.766	0.784
	Correction 1	0.843	0.856	0.720	0.817	0.712	0.828	0.767	0.781
	Correction 2	0.843	0.855	0.719	0.810	0.712	0.824	0.767	0.779
	Correction 3	0.843	0.844	0.726	0.841	0.715	0.818	0.766	0.768
Overall SOFA score increase	Reference	0.854	0.880	0.754	0.931	0.609	0.700	0.776	0.719
	SpO2	0.854	0.872	0.756	0.843	0.607	0.713	0.774	0.761
	Correction 1	0.854	0.882	0.756	0.814	0.609	0.728	0.776	0.783
	Correction 2	0.855	0.882	0.756	0.817	0.609	0.732	0.776	0.785
	Correction 3	0.855	0.862	0.758	0.840	0.610	0.705	0.775	0.753

Table B.4: Fairness metrics in the training set, in the Blood-gas and Oximetry study.

Task	Feature	**LR**		**RF**		**XGBoost**	
		Demographic parity difference	Equalized odds ratio	Demographic parity difference	Equalized odds ratio	Demographic parity difference	Equalized odds ratio
In-hospital mortality	Reference	0.042	0.857	0.068	1.000	0.080	0.754
	SpO2	0.046	0.840	0.068	1.000	0.080	0.756
	Correction 1	0.041	0.852	0.068	1.000	0.076	0.764
	Correction 2	0.036	0.849	0.068	1.000	0.077	0.764
	Correction 3	0.042	0.853	0.068	1.000	0.080	0.755
Respiratory SOFA score	Reference	0.044	0.830	0.060	0.000	0.032	0.833
	SpO2	0.045	0.839	0.060	0.000	0.045	0.786
	Correction 1	0.043	0.849	0.060	0.000	0.031	0.833
	Correction 2	0.043	0.853	0.060	0.000	0.030	0.828
	Correction 3	0.041	0.845	0.060	0.000	0.029	0.834
Overall SOFA score increase	Reference	0.059	0.816	0.022	0.000	0.039	0.849
	SpO2	0.056	0.822	0.022	0.000	0.036	0.868
	Correction 1	0.060	0.814	0.022	0.000	0.042	0.857
	Correction 2	0.058	0.823	0.022	0.000	0.042	0.846
	Correction 3	0.061	0.809	0.022	0.000	0.039	0.856

B.1.2 Test set

Table B.5: AUROC and recall in the test set with LR, in the Blood-gas and Oximetry study.

Task	Feature		AUROC						Recall					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.708	0.721	0.709	0.745	0.729	0.704	0.621	0.617	0.604	0.655	0.66	0.619
	SpO2		0.708	0.722	0.709	0.743	0.727	0.704	0.622	0.599	0.6	0.653	0.662	0.621
		P-value	0.496	0.604	0.768	0.488	0.186	1	0.825	0.056	0.618	0.786	0.78	0.443
		95% CI	(-0.001 - 0.001)	(-0.007 - 0.004)	(-0.005 - 0.004)	(-0.003 - 0.007)	(-0.001 - 0.005)	(-0.001 - 0.001)	(-0.007 - 0.005)	(-0.001 - 0.036)	(-0.012 - 0.019)	(-0.012 - 0.015)	(-0.017 - 0.013)	(-0.009 - 0.004)
	C1		0.708	0.729	0.706	0.739	0.73	0.703	0.621	0.619	0.599	0.645	0.666	0.619
		P-value	0.475	0.012	0.134	0.201	0.785	0.637	0.904	0.83	0.241	0.518	0.539	0.786
		95% CI	(-0.001 - 0.002)	(-0.015 - -0.002)	(-0.001 - 0.006)	(-0.004 - 0.015)	(-0.006 - 0.005)	(-0.001 - 0.002)	(-0.003 - 0.004)	(-0.024 - 0.019)	(-0.004 - 0.013)	(-0.022 - 0.041)	(-0.027 - 0.015)	(-0.004 - 0.003)
	C2		0.708	0.728	0.708	0.742	0.729	0.704	0.622	0.604	0.607	0.657	0.662	0.618
		P-value	1	0.123	0.71	0.139	0.952	1	0.971	0.213	0.623	0.693	0.785	0.843
		95% CI	(-0.001 - 0.001)	(-0.017 - 0.002)	(-0.003 - 0.005)	(-0.001 - 0.007)	(-0.004 - 0.004)	(-0.001 - 0.001)	(-0.006 - 0.006)	(-0.009 - 0.034)	(-0.016 - 0.01)	(-0.018 - 0.012)	(-0.015 - 0.011)	(-0.005 - 0.006)
	C3		0.709	0.73	0.709	0.742	0.732	0.705	0.619	0.622	0.594	0.648	0.661	0.618
		P-value	0.068	0.081	0.605	0.305	0.18	0.068	0.303	0.558	0.082	0.367	0.803	0.666
		95% CI	(-0.002 - 0.0)	(-0.02 - 0.001)	(-0.003 - 0.002)	(-0.003 - 0.01)	(-0.008 - 0.002)	(-0.002 - 0.0)	(-0.002 - 0.006)	(-0.023 - 0.013)	(-0.002 - 0.022)	(-0.009 - 0.023)	(-0.012 - 0.009)	(-0.003 - 0.004)
Respiratory SOFA score	R		0.758	0.788	0.772	0.763	0.768	0.752	0.692	0.704	0.691	0.7	0.693	0.689
	SpO2		0.754	0.782	0.765	0.761	0.765	0.749	0.688	0.7	0.686	0.695	0.689	0.684
		P-value	<.001	0.031	<.01	0.402	0.106	<.001	<.01	0.396	0.213	0.353	0.05	<.01
		95% CI	(0.002 - 0.005)	(0.001 - 0.013)	(0.003 - 0.011)	(-0.002 - 0.004)	(-0.001 - 0.007)	(0.002 - 0.005)	(0.002 - 0.007)	(-0.006 - 0.014)	(-0.004 - 0.014)	(-0.007 - 0.018)	(0.0 - 0.009)	(0.001 - 0.008)
	C1		0.754	0.784	0.769	0.757	0.763	0.748	0.689	0.703	0.694	0.697	0.687	0.684
		P-value	<.01	<.01	0.169	<.001	<.01	<.01	0.015	0.847	0.255	0.532	0.178	<.01
		95% CI	(0.002 - 0.006)	(0.002 - 0.008)	(-0.001 - 0.007)	(0.003 - 0.008)	(0.002 - 0.009)	(0.002 - 0.006)	(0.001 - 0.006)	(-0.016 - 0.019)	(-0.009 - 0.003)	(-0.008 - 0.014)	(-0.003 - 0.016)	(0.002 - 0.007)
	C2		0.755	0.784	0.768	0.761	0.764	0.749	0.69	0.698	0.692	0.701	0.69	0.685
		P-value	0.02	0.13	0.239	0.194	0.017	0.017	0.043	0.489	0.543	0.966	0.404	0.012
		95% CI	(0.001 - 0.005)	(-0.002 - 0.011)	(-0.003 - 0.009)	(-0.001 - 0.004)	(0.001 - 0.007)	(0.001 - 0.005)	(0.0 - 0.005)	(-0.012 - 0.024)	(-0.006 - 0.004)	(-0.011 - 0.01)	(-0.006 - 0.013)	(0.001 - 0.006)
	C3		0.753	0.779	0.769	0.758	0.762	0.748	0.688	0.698	0.694	0.692	0.688	0.684
		P-value	<.001	<.01	0.226	0.029	<.001	<.001	0.017	0.597	0.496	0.153	0.14	<.01
		95% CI	(0.004 - 0.005)	(0.005 - 0.014)	(-0.002 - 0.008)	(0.001 - 0.008)	(0.004 - 0.009)	(0.003 - 0.006)	(0.001 - 0.008)	(-0.018 - 0.03)	(-0.011 - 0.006)	(-0.004 - 0.022)	(-0.002 - 0.012)	(0.003 - 0.008)
Overall SOFA score increase	R		0.732	0.708	0.724	0.725	0.735	0.734	0.679	0.684	0.682	0.662	0.677	0.68
	SpO2		0.731	0.708	0.723	0.722	0.734	0.733	0.678	0.679	0.675	0.66	0.675	0.68
		P-value	0.052	0.965	0.271	<.01	0.648	0.313	0.407	0.55	0.017	0.369	0.519	0.884
		95% CI	(-0.0 - 0.002)	(-0.005 - 0.005)	(-0.001 - 0.003)	(0.001 - 0.005)	(-0.002 - 0.002)	(-0.001 - 0.002)	(-0.002 - 0.004)	(-0.013 - 0.023)	(0.002 - 0.014)	(-0.004 - 0.009)	(-0.005 - 0.01)	(-0.003 - 0.003)
	C1		0.73	0.706	0.724	0.722	0.734	0.733	0.677	0.678	0.681	0.663	0.674	0.678
		P-value	<.01	0.276	0.705	0.083	0.596	0.093	0.184	0.484	0.751	0.828	0.493	0.259
		95% CI	(0.0 - 0.002)	(-0.002 - 0.007)	(-0.002 - 0.003)	(-0.001 - 0.007)	(-0.002 - 0.003)	(-0.0 - 0.002)	(-0.001 - 0.005)	(-0.013 - 0.025)	(-0.006 - 0.009)	(-0.01 - 0.008)	(-0.007 - 0.013)	(-0.002 - 0.006)
	C2		0.731	0.708	0.724	0.722	0.734	0.733	0.678	0.681	0.682	0.663	0.676	0.679
		P-value	0.017	0.883	0.625	0.057	0.531	0.081	0.554	0.797	0.841	0.916	0.756	0.498
		95% CI	(0.0 - 0.002)	(-0.004 - 0.005)	(-0.001 - 0.002)	(-0.0 - 0.007)	(-0.001 - 0.003)	(-0.0 - 0.002)	(-0.002 - 0.004)	(-0.02 - 0.026)	(-0.007 - 0.008)	(-0.009 - 0.008)	(-0.007 - 0.01)	(-0.002 - 0.005)
	C3		0.73	0.706	0.723	0.721	0.734	0.732	0.677	0.67	0.683	0.659	0.676	0.679
		P-value	<.01	0.295	0.435	0.025	0.223	0.012	0.402	0.188	0.894	0.604	0.481	0.413
		95% CI	(0.001 - 0.003)	(-0.003 - 0.008)	(-0.001 - 0.003)	(0.001 - 0.008)	(-0.001 - 0.002)	(0.0 - 0.003)	(-0.002 - 0.005)	(-0.008 - 0.036)	(-0.007 - 0.006)	(-0.011 - 0.018)	(-0.003 - 0.005)	(-0.002 - 0.004)

Table B.6: F1-score and accuracy in the test set with LR, in the Blood-gas and Oximetry study.

Task	Feature		F1-score						Accuracy					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.474	0.527	0.47	0.545	0.475	0.466	0.669	0.685	0.679	0.691	0.675	0.665
	SpO2		0.472	0.518	0.467	0.54	0.472	0.465	0.666	0.682	0.678	0.686	0.671	0.662
		P-value	<.01	0.158	0.506	0.404	0.328	0.067	0.044	0.693	0.825	0.303	0.113	0.029
		95% CI	(0.001 - 0.003)	(-0.004 - 0.022)	(-0.006 - 0.012)	(-0.007 - 0.015)	(-0.003 - 0.009)	(-0.0 - 0.003)	(0.0 - 0.006)	(-0.011 - 0.016)	(-0.004 - 0.005)	(-0.005 - 0.015)	(-0.001 - 0.01)	(0.0 - 0.005)
	C1		0.473	0.53	0.466	0.537	0.476	0.466	0.668	0.691	0.676	0.687	0.674	0.665
		P-value	0.509	0.718	0.031	0.471	0.827	0.825	0.555	0.471	0.192	0.491	0.487	0.922
		95% CI	(-0.001 - 0.003)	(-0.023 - 0.016)	(0.0 - 0.008)	(-0.014 - 0.029)	(-0.009 - 0.007)	(-0.002 - 0.002)	(-0.002 - 0.003)	(-0.023 - 0.011)	(-0.002 - 0.008)	(-0.007 - 0.014)	(-0.003 - 0.007)	(-0.002 - 0.002)
	C2		0.473	0.532	0.47	0.545	0.476	0.465	0.668	0.699	0.677	0.689	0.676	0.664
		P-value	0.637	0.455	0.953	0.958	0.833	0.475	0.623	0.047	0.305	0.734	0.76	0.597
		95% CI	(-0.002 - 0.003)	(-0.023 - 0.011)	(-0.007 - 0.008)	(-0.008 - 0.008)	(-0.006 - 0.005)	(-0.002 - 0.004)	(-0.002 - 0.004)	(-0.028 - -0.0)	(-0.002 - 0.006)	(-0.007 - 0.01)	(-0.007 - 0.005)	(-0.003 - 0.005)
	C3		0.474	0.534	0.466	0.542	0.477	0.466	0.67	0.693	0.679	0.691	0.677	0.666
		P-value	0.916	0.319	0.204	0.618	0.662	0.771	0.17	0.34	0.683	0.821	0.631	0.409
		95% CI	(-0.002 - 0.002)	(-0.025 - 0.009)	(-0.003 - 0.012)	(-0.008 - 0.012)	(-0.009 - 0.006)	(-0.003 - 0.002)	(-0.003 - 0.001)	(-0.025 - 0.01)	(-0.004 - 0.003)	(-0.006 - 0.005)	(-0.008 - 0.005)	(-0.004 - 0.002)
Respiratory SOFA score	R		0.659	0.665	0.647	0.672	0.666	0.655	0.701	0.737	0.718	0.705	0.711	0.697
	SpO2		0.657	0.664	0.646	0.672	0.666	0.653	0.701	0.737	0.719	0.706	0.713	0.696
		P-value	<.01	0.865	0.616	0.856	0.819	0.013	0.509	0.925	0.578	0.734	0.127	0.137
		95% CI	(0.001 - 0.003)	(-0.01 - 0.011)	(-0.003 - 0.005)	(-0.007 - 0.008)	(-0.003 - 0.003)	(0.001 - 0.003)	(-0.0 - 0.001)	(-0.01 - 0.009)	(-0.003 - 0.002)	(-0.007 - 0.005)	(-0.004 - 0.001)	(-0.0 - 0.002)
	C1		0.656	0.662	0.649	0.669	0.661	0.653	0.7	0.735	0.719	0.701	0.708	0.695
		P-value	<.01	0.538	0.379	0.272	0.134	0.019	0.03	0.618	0.561	0.212	0.177	0.045
		95% CI	(0.001 - 0.004)	(-0.008 - 0.014)	(-0.008 - 0.003)	(-0.004 - 0.011)	(-0.002 - 0.01)	(0.0 - 0.004)	(0.0 - 0.002)	(-0.006 - 0.01)	(-0.006 - 0.003)	(-0.002 - 0.009)	(-0.002 - 0.008)	(0.0 - 0.003)
	C2		0.657	0.663	0.648	0.672	0.662	0.653	0.701	0.738	0.718	0.704	0.708	0.695
		P-value	0.035	0.795	0.825	0.802	0.203	0.06	0.322	0.815	0.85	0.683	0.228	0.161
		95% CI	(0.0 - 0.003)	(-0.012 - 0.015)	(-0.005 - 0.004)	(-0.005 - 0.007)	(-0.002 - 0.009)	(-0.0 - 0.004)	(-0.001 - 0.002)	(-0.012 - 0.01)	(-0.004 - 0.005)	(-0.003 - 0.005)	(-0.002 - 0.008)	(-0.001 - 0.003)
	C3		0.657	0.66	0.649	0.669	0.662	0.653	0.701	0.735	0.718	0.704	0.71	0.696
		P-value	<.01	0.516	0.536	0.346	0.269	<.01	0.413	0.722	0.767	0.759	0.493	0.405
		95% CI	(0.001 - 0.004)	(-0.011 - 0.02)	(-0.008 - 0.004)	(-0.004 - 0.01)	(-0.003 - 0.009)	(0.001 - 0.004)	(-0.001 - 0.002)	(-0.009 - 0.012)	(-0.005 - 0.004)	(-0.004 - 0.006)	(-0.003 - 0.007)	(-0.001 - 0.002)
Overall SOFA score increase	R		0.491	0.453	0.483	0.481	0.502	0.493	0.665	0.622	0.658	0.671	0.667	0.667
	SpO2		0.49	0.451	0.482	0.477	0.503	0.491	0.664	0.623	0.659	0.666	0.669	0.665
		P-value	0.07	0.602	0.343	0.08	0.481	0.2	0.168	0.812	0.371	0.057	0.086	0.068
		95% CI	(-0.0 - 0.003)	(-0.006 - 0.01)	(-0.002 - 0.005)	(-0.001 - 0.01)	(-0.004 - 0.002)	(-0.001 - 0.003)	(-0.001 - 0.003)	(-0.01 - 0.008)	(-0.005 - 0.002)	(-0.0 - 0.01)	(-0.005 - 0.0)	(-0.0 - 0.004)
	C1		0.489	0.445	0.481	0.478	0.5	0.491	0.663	0.613	0.655	0.667	0.665	0.665
		P-value	0.023	0.049	0.311	0.359	0.29	0.164	0.054	<.01	0.221	0.11	0.353	0.153
		95% CI	(0.0 - 0.004)	(0.0 - 0.016)	(-0.002 - 0.007)	(-0.004 - 0.009)	(-0.003 - 0.008)	(-0.001 - 0.004)	(-0.0 - 0.004)	(0.004 - 0.015)	(-0.002 - 0.007)	(-0.001 - 0.009)	(-0.002 - 0.005)	(-0.001 - 0.004)
	C2		0.49	0.449	0.482	0.478	0.502	0.492	0.664	0.618	0.656	0.667	0.667	0.666
		P-value	0.269	0.459	0.492	0.224	0.962	0.569	0.309	0.195	0.264	0.091	0.766	0.515
		95% CI	(-0.001 - 0.002)	(-0.007 - 0.015)	(-0.003 - 0.005)	(-0.002 - 0.009)	(-0.004 - 0.005)	(-0.001 - 0.002)	(-0.001 - 0.002)	(-0.003 - 0.012)	(-0.002 - 0.005)	(-0.001 - 0.01)	(-0.004 - 0.003)	(-0.001 - 0.003)
	C3		0.489	0.44	0.482	0.476	0.502	0.491	0.663	0.61	0.655	0.666	0.667	0.665
		P-value	<.01	0.012	0.293	0.158	0.95	0.06	0.042	<.01	0.161	0.056	0.656	0.123
		95% CI	(0.001 - 0.003)	(0.004 - 0.022)	(-0.002 - 0.005)	(-0.003 - 0.014)	(-0.003 - 0.004)	(-0.0 - 0.003)	(0.0 - 0.004)	(0.005 - 0.02)	(-0.001 - 0.006)	(-0.0 - 0.011)	(-0.004 - 0.003)	(-0.001 - 0.005)

Table B.7: AUROC and recall in the test set with RF, in the Blood-gas and Oximetry study.

Task	Feature		AUROC						Recall					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.734	0.753	0.73	0.762	0.757	0.731	0.114	0.115	0.116	0.124	0.146	0.11
	SpO2		0.735	0.749	0.731	0.76	0.758	0.732	0.117	0.137	0.121	0.122	0.159	0.113
		P-value	0.482	0.651	0.769	0.515	0.68	0.351	0.065	0.174	0.407	0.747	0.06	0.202
		95% CI	(-0.003 - 0.001)	(-0.014 - 0.021)	(-0.009 - 0.007)	(-0.004 - 0.008)	(-0.008 - 0.006)	(-0.003 - 0.001)	(-0.008 - 0.0)	(-0.057 - 0.012)	(-0.017 - 0.007)	(-0.013 - 0.017)	(-0.027 - 0.001)	(-0.009 - 0.002)
	C1		0.734	0.751	0.726	0.762	0.756	0.731	0.114	0.12	0.119	0.128	0.145	0.11
		P-value	0.924	0.81	0.367	0.948	0.879	0.874	0.748	0.782	0.538	0.747	0.846	1
		95% CI	(-0.002 - 0.002)	(-0.017 - 0.021)	(-0.004 - 0.01)	(-0.011 - 0.01)	(-0.008 - 0.009)	(-0.003 - 0.003)	(-0.005 - 0.004)	(-0.048 - 0.037)	(-0.012 - 0.007)	(-0.027 - 0.02)	(-0.012 - 0.015)	(-0.004 - 0.004)
	C2		0.734	0.76	0.728	0.756	0.76	0.731	0.115	0.127	0.126	0.13	0.144	0.11
		P-value	1	0.302	0.545	0.158	0.432	0.876	0.289	0.521	0.167	0.484	0.806	0.759
		95% CI	(-0.002 - 0.002)	(-0.024 - 0.008)	(-0.004 - 0.008)	(-0.003 - 0.014)	(-0.011 - 0.005)	(-0.002 - 0.001)	(-0.005 - 0.002)	(-0.054 - 0.029)	(-0.025 - 0.005)	(-0.026 - 0.013)	(-0.017 - 0.021)	(-0.005 - 0.004)
	C3		0.736	0.762	0.728	0.759	0.758	0.732	0.119	0.138	0.125	0.129	0.157	0.114
		P-value	0.186	0.135	0.759	0.678	0.856	0.315	0.015	0.225	0.092	0.31	0.104	0.153
		95% CI	(-0.003 - 0.001)	(-0.022 - 0.004)	(-0.007 - 0.009)	(-0.01 - 0.014)	(-0.008 - 0.007)	(-0.004 - 0.001)	(-0.009 - -0.001)	(-0.063 - 0.017)	(-0.02 - 0.002)	(-0.015 - 0.005)	(-0.024 - 0.003)	(-0.009 - 0.002)
Respiratory SOFA score	R		0.775	0.8	0.788	0.777	0.787	0.769	0.556	0.549	0.562	0.516	0.578	0.552
	SpO2		0.769	0.797	0.778	0.774	0.785	0.763	0.551	0.542	0.53	0.502	0.581	0.55
		P-value	<.001	0.493	<.001	0.133	0.251	<.001	0.022	0.569	<.001	0.043	0.558	0.333
		95% CI	(0.004 - 0.007)	(-0.006 - 0.012)	(0.005 - 0.013)	(-0.001 - 0.009)	(-0.002 - 0.007)	(0.004 - 0.008)	(0.001 - 0.01)	(-0.02 - 0.035)	(0.019 - 0.047)	(0.001 - 0.027)	(-0.014 - 0.008)	(-0.002 - 0.006)
	C1		0.769	0.794	0.782	0.77	0.783	0.763	0.55	0.56	0.559	0.498	0.58	0.545
		P-value	<.001	0.244	0.06	0.034	0.11	<.001	<.01	0.321	0.505	0.071	0.702	<.01
		95% CI	(0.004 - 0.008)	(-0.004 - 0.015)	(-0.0 - 0.012)	(0.001 - 0.013)	(-0.001 - 0.008)	(0.004 - 0.008)	(0.002 - 0.01)	(-0.036 - 0.013)	(-0.007 - 0.014)	(-0.002 - 0.036)	(-0.017 - 0.012)	(0.002 - 0.012)
	C2		0.768	0.791	0.782	0.77	0.783	0.762	0.549	0.55	0.556	0.5	0.581	0.544
		P-value	<.001	0.035	0.02	0.016	<.01	<.001	<.01	0.962	0.313	0.047	0.641	0.016
		95% CI	(0.005 - 0.009)	(0.001 - 0.017)	(0.001 - 0.01)	(0.002 - 0.013)	(0.001 - 0.007)	(0.005 - 0.009)	(0.003 - 0.011)	(-0.024 - 0.023)	(-0.007 - 0.019)	(0.0 - 0.032)	(-0.017 - 0.011)	(0.002 - 0.014)
	C3		0.768	0.79	0.781	0.766	0.783	0.762	0.55	0.546	0.557	0.478	0.581	0.547
		P-value	<.001	0.092	<.01	0.014	0.104	<.001	<.01	0.871	0.438	0.012	0.429	0.022
		95% CI	(0.005 - 0.009)	(-0.002 - 0.021)	(0.002 - 0.011)	(0.003 - 0.02)	(-0.001 - 0.009)	(0.005 - 0.009)	(0.003 - 0.009)	(-0.035 - 0.041)	(-0.009 - 0.019)	(0.011 - 0.065)	(-0.012 - 0.006)	(0.001 - 0.01)
Overall SOFA score increase	R		0.752	0.747	0.74	0.753	0.752	0.755	0.176	0.152	0.17	0.123	0.181	0.18
	SpO2		0.751	0.743	0.743	0.75	0.753	0.752	0.173	0.155	0.171	0.129	0.182	0.175
		P-value	0.096	0.467	0.327	0.278	0.871	0.104	0.102	0.831	0.739	0.259	0.853	0.059
		95% CI	(-0.0 - 0.003)	(-0.008 - 0.016)	(-0.009 - 0.003)	(-0.003 - 0.01)	(-0.007 - 0.006)	(-0.001 - 0.005)	(-0.001 - 0.007)	(-0.025 - 0.02)	(-0.012 - 0.009)	(-0.018 - 0.006)	(-0.013 - 0.011)	(-0.0 - 0.01)
	C1		0.751	0.736	0.743	0.744	0.756	0.753	0.173	0.155	0.168	0.126	0.183	0.177
		P-value	0.134	0.122	0.111	0.029	0.25	0.117	0.181	0.775	0.693	0.64	0.659	0.176
		95% CI	(-0.001 - 0.004)	(-0.003 - 0.025)	(-0.007 - 0.001)	(0.001 - 0.016)	(-0.01 - 0.003)	(-0.001 - 0.005)	(-0.001 - 0.006)	(-0.021 - 0.016)	(-0.006 - 0.009)	(-0.019 - 0.013)	(-0.012 - 0.008)	(-0.002 - 0.008)
	C2		0.75	0.734	0.743	0.748	0.757	0.751	0.173	0.144	0.175	0.13	0.178	0.176
		P-value	<.01	<.01	0.383	0.213	0.118	<.01	0.321	0.529	0.187	0.325	0.748	0.115
		95% CI	(0.001 - 0.003)	(0.004 - 0.023)	(-0.01 - 0.004)	(-0.003 - 0.013)	(-0.011 - 0.002)	(0.002 - 0.005)	(-0.003 - 0.008)	(-0.02 - 0.037)	(-0.013 - 0.003)	(-0.021 - 0.008)	(-0.015 - 0.02)	(-0.001 - 0.01)
	C3		0.75	0.739	0.742	0.746	0.754	0.751	0.173	0.146	0.174	0.126	0.179	0.176
		P-value	<.01	0.135	0.686	0.1	0.546	0.019	0.059	0.465	0.331	0.551	0.791	0.042
		95% CI	(0.001 - 0.004)	(-0.003 - 0.02)	(-0.008 - 0.006)	(-0.002 - 0.016)	(-0.009 - 0.005)	(0.001 - 0.007)	(-0.0 - 0.006)	(-0.013 - 0.027)	(-0.014 - 0.005)	(-0.014 - 0.008)	(-0.015 - 0.019)	(0.0 - 0.008)

Table B.8: F1-score and accuracy in the test set with RF, in the Blood-gas and Oximetry study.

Task	Feature		F1-score						Accuracy					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.192	0.185	0.191	0.204	0.238	0.186	0.771	0.728	0.769	0.732	0.793	0.773
	SpO2		0.198	0.223	0.199	0.2	0.255	0.191	0.772	0.735	0.772	0.731	0.795	0.774
		P-value	0.042	0.152	0.302	0.639	0.063	0.196	0.074	0.259	0.02	0.464	0.436	0.382
		95% CI	(-0.012 - -0.0)	(-0.091 - 0.017)	(-0.024 - 0.008)	(-0.017 - 0.027)	(-0.036 - 0.001)	(-0.014 - 0.003)	(-0.002 - 0.0)	(-0.02 - 0.006)	(-0.006 - -0.001)	(-0.003 - 0.006)	(-0.006 - 0.003)	(-0.002 - 0.001)
	C1		0.193	0.203	0.196	0.208	0.236	0.186	0.771	0.735	0.771	0.734	0.792	0.773
		P-value	0.669	0.52	0.389	0.782	0.776	0.948	0.685	0.305	0.042	0.551	0.419	1
		95% CI	(-0.007 - 0.005)	(-0.078 - 0.042)	(-0.018 - 0.008)	(-0.036 - 0.028)	(-0.016 - 0.021)	(-0.007 - 0.007)	(-0.002 - 0.001)	(-0.022 - 0.008)	(-0.005 - -0.0)	(-0.008 - 0.005)	(-0.002 - 0.005)	(-0.002 - 0.002)
	C2		0.195	0.211	0.206	0.211	0.235	0.187	0.772	0.733	0.772	0.734	0.792	0.773
		P-value	0.151	0.348	0.142	0.589	0.797	0.687	0.153	0.463	0.069	0.585	0.69	0.872
		95% CI	(-0.009 - 0.002)	(-0.083 - 0.032)	(-0.036 - 0.006)	(-0.033 - 0.02)	(-0.023 - 0.03)	(-0.008 - 0.005)	(-0.002 - 0.0)	(-0.021 - 0.01)	(-0.007 - 0.0)	(-0.006 - 0.004)	(-0.004 - 0.005)	(-0.001 - 0.001)
	C3		0.2	0.228	0.205	0.213	0.254	0.191	0.772	0.738	0.772	0.736	0.795	0.773
		P-value	0.017	0.124	0.077	0.212	0.092	0.186	0.223	0.172	0.06	0.024	0.238	0.893
		95% CI	(-0.014 - -0.002)	(-0.1 - 0.014)	(-0.03 - 0.002)	(-0.023 - 0.006)	(-0.035 - 0.003)	(-0.013 - 0.003)	(-0.002 - 0.001)	(-0.026 - 0.005)	(-0.006 - 0.0)	(-0.007 - -0.001)	(-0.006 - 0.002)	(-0.002 - 0.002)
Respiratory SOFA score	R		0.626	0.618	0.625	0.611	0.642	0.622	0.724	0.75	0.748	0.716	0.732	0.72
	SpO2		0.62	0.615	0.602	0.599	0.643	0.617	0.719	0.75	0.738	0.711	0.732	0.714
		P-value	<.01	0.793	<.01	0.063	0.863	<.01	<.01	0.984	0.011	0.15	1	<.001
		95% CI	(0.003 - 0.01)	(-0.018 - 0.023)	(0.011 - 0.034)	(-0.001 - 0.024)	(-0.012 - 0.011)	(0.002 - 0.008)	(0.002 - 0.007)	(-0.011 - 0.011)	(0.003 - 0.016)	(-0.002 - 0.013)	(-0.009 - 0.009)	(0.003 - 0.007)
	C1		0.618	0.625	0.618	0.594	0.645	0.613	0.718	0.75	0.742	0.707	0.735	0.712
		P-value	<.001	0.425	0.096	0.021	0.6	<.001	<.001	0.953	0.03	<.01	0.485	<.001
		95% CI	(0.005 - 0.011)	(-0.027 - 0.013)	(-0.001 - 0.015)	(0.003 - 0.031)	(-0.015 - 0.009)	(0.005 - 0.013)	(0.004 - 0.009)	(-0.012 - 0.011)	(0.001 - 0.01)	(0.004 - 0.015)	(-0.011 - 0.005)	(0.005 - 0.01)
	C2		0.617	0.615	0.615	0.596	0.642	0.613	0.717	0.747	0.74	0.708	0.731	0.712
		P-value	<.001	0.779	0.04	0.015	0.963	<.01	<.001	0.56	<.01	0.032	0.756	<.001
		95% CI	(0.005 - 0.012)	(-0.016 - 0.02)	(0.001 - 0.019)	(0.004 - 0.026)	(-0.01 - 0.009)	(0.005 - 0.014)	(0.004 - 0.009)	(-0.007 - 0.012)	(0.003 - 0.012)	(0.001 - 0.015)	(-0.005 - 0.006)	(0.004 - 0.01)
	C3		0.619	0.618	0.618	0.579	0.644	0.615	0.718	0.752	0.743	0.703	0.734	0.713
		P-value	<.001	1	0.19	<.01	0.506	<.01	<.01	0.843	0.087	0.011	0.638	<.01
		95% CI	(0.004 - 0.01)	(-0.032 - 0.032)	(-0.004 - 0.017)	(0.01 - 0.052)	(-0.01 - 0.005)	(0.003 - 0.011)	(0.003 - 0.008)	(-0.019 - 0.016)	(-0.001 - 0.01)	(0.004 - 0.023)	(-0.007 - 0.004)	(0.003 - 0.009)
Overall SOFA score increase	R		0.28	0.242	0.268	0.205	0.286	0.286	0.785	0.783	0.782	0.781	0.777	0.786
	SpO2		0.276	0.249	0.272	0.216	0.289	0.28	0.785	0.788	0.785	0.785	0.778	0.785
		P-value	0.128	0.66	0.591	0.225	0.684	0.063	1	0.199	0.148	0.066	0.382	0.107
		95% CI	(-0.001 - 0.009)	(-0.04 - 0.027)	(-0.018 - 0.011)	(-0.03 - 0.008)	(-0.019 - 0.013)	(-0.0 - 0.013)	(-0.001 - 0.001)	(-0.013 - 0.003)	(-0.005 - 0.001)	(-0.007 - 0.0)	(-0.005 - 0.002)	(-0.0 - 0.002)
	C1		0.277	0.246	0.268	0.212	0.29	0.282	0.785	0.784	0.784	0.784	0.778	0.785
		P-value	0.235	0.702	0.937	0.521	0.557	0.178	1	0.76	0.28	0.111	0.545	0.193
		95% CI	(-0.002 - 0.007)	(-0.029 - 0.02)	(-0.011 - 0.012)	(-0.03 - 0.016)	(-0.018 - 0.011)	(-0.002 - 0.011)	(-0.001 - 0.001)	(-0.007 - 0.005)	(-0.005 - 0.002)	(-0.007 - 0.001)	(-0.005 - 0.003)	(-0.001 - 0.002)
	C2		0.276	0.228	0.276	0.216	0.283	0.281	0.784	0.78	0.785	0.784	0.776	0.785
		P-value	0.268	0.46	0.146	0.318	0.748	0.146	0.394	0.466	0.067	0.272	0.728	0.193
		95% CI	(-0.003 - 0.011)	(-0.028 - 0.057)	(-0.019 - 0.003)	(-0.033 - 0.012)	(-0.02 - 0.027)	(-0.002 - 0.014)	(-0.001 - 0.002)	(-0.006 - 0.012)	(-0.005 - 0.0)	(-0.007 - 0.002)	(-0.005 - 0.007)	(-0.001 - 0.003)
	C3		0.276	0.23	0.275	0.212	0.282	0.28	0.784	0.779	0.785	0.784	0.774	0.785
		P-value	0.054	0.386	0.237	0.462	0.664	0.044	0.117	0.236	0.056	0.147	0.265	0.042
		95% CI	(-0.0 - 0.008)	(-0.017 - 0.041)	(-0.02 - 0.006)	(-0.024 - 0.012)	(-0.019 - 0.028)	(0.0 - 0.011)	(-0.0 - 0.002)	(-0.003 - 0.011)	(-0.005 - 0.0)	(-0.007 - 0.001)	(-0.003 - 0.008)	(0.0 - 0.003)

Table B.9: AUROC and recall in the test set with XGBoost, in the Blood-gas and Oximetry study.

Task	Feature		AUROC						Recall					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.728	0.762	0.717	0.751	0.754	0.725	0.581	0.635	0.549	0.64	0.612	0.577
	SpO2	P-value	0.726	0.739	0.714	0.749	0.74	0.725	0.577	0.588	0.533	0.644	0.581	0.578
		95% CI	0.414	0.066	0.594	0.697	<.01	0.92	0.207	0.085	0.112	0.705	<.01	0.829
			(-0.003 - 0.006)	(-0.002 - 0.049)	(-0.008 - 0.013)	(-0.009 - 0.013)	(0.006 - 0.023)	(-0.007 - 0.006)	(-0.003 - 0.012)	(-0.008 - 0.102)	(-0.005 - 0.038)	(-0.028 - 0.02)	(0.01 - 0.051)	(-0.011 - 0.009)
	C1	P-value	0.727	0.764	0.717	0.747	0.748	0.724	0.578	0.614	0.549	0.641	0.598	0.575
		95% CI	0.721	0.886	0.938	0.494	0.304	0.72	0.485	0.465	0.95	0.904	0.213	0.505
			(-0.006 - 0.008)	(-0.029 - 0.026)	(-0.012 - 0.011)	(-0.01 - 0.019)	(-0.006 - 0.018)	(-0.005 - 0.007)	(-0.007 - 0.013)	(-0.042 - 0.085)	(-0.024 - 0.025)	(-0.029 - 0.026)	(-0.009 - 0.037)	(-0.006 - 0.012)
	C2	P-value	0.729	0.775	0.714	0.744	0.747	0.726	0.582	0.635	0.548	0.64	0.599	0.579
		95% CI	0.781	0.191	0.69	0.343	0.23	0.541	0.804	0.974	0.908	0.995	0.102	0.66
			(-0.005 - 0.004)	(-0.033 - 0.008)	(-0.01 - 0.015)	(-0.01 - 0.025)	(-0.005 - 0.019)	(-0.007 - 0.004)	(-0.008 - 0.006)	(-0.04 - 0.042)	(-0.018 - 0.02)	(-0.033 - 0.033)	(-0.003 - 0.03)	(-0.008 - 0.005)
	C3	P-value	0.729	0.779	0.717	0.748	0.746	0.726	0.581	0.654	0.549	0.629	0.59	0.579
		95% CI	0.764	0.226	0.79	0.598	0.193	0.657	0.836	0.421	0.892	0.425	0.065	0.78
			(-0.006 - 0.004)	(-0.046 - 0.012)	(-0.008 - 0.007)	(-0.01 - 0.017)	(-0.005 - 0.022)	(-0.008 - 0.005)	(-0.007 - 0.008)	(-0.071 - 0.033)	(-0.012 - 0.014)	(-0.018 - 0.038)	(-0.002 - 0.046)	(-0.011 - 0.008)
Respiratory SOFA score	R		0.774	0.802	0.785	0.773	0.788	0.768	0.664	0.683	0.672	0.64	0.686	0.66
	SpO2	P-value	0.769	0.788	0.78	0.766	0.784	0.764	0.662	0.659	0.652	0.63	0.682	0.66
		95% CI	<.01	<.01	0.213	0.117	0.209	<.01	0.238	0.069	<.01	0.126	0.422	0.93
			(0.002 - 0.007)	(0.005 - 0.023)	(-0.003 - 0.012)	(-0.002 - 0.016)	(-0.003 - 0.012)	(0.002 - 0.007)	(-0.002 - 0.006)	(-0.002 - 0.051)	(0.008 - 0.03)	(-0.003 - 0.023)	(-0.007 - 0.015)	(-0.005 - 0.005)
	C1	P-value	0.769	0.79	0.781	0.763	0.784	0.763	0.664	0.666	0.676	0.625	0.679	0.66
		95% CI	<.001	0.047	0.15	0.037	0.118	<.001	1	0.229	0.319	0.033	0.238	0.636
			(0.003 - 0.007)	(0.0 - 0.023)	(-0.002 - 0.01)	(0.001 - 0.019)	(-0.001 - 0.009)	(0.003 - 0.007)	(-0.004 - 0.004)	(-0.013 - 0.047)	(-0.015 - 0.005)	(0.002 - 0.028)	(-0.005 - 0.018)	(-0.004 - 0.003)
	C2	P-value	0.768	0.785	0.779	0.763	0.784	0.763	0.66	0.66	0.667	0.624	0.687	0.656
		95% CI	<.001	0.012	0.072	0.019	0.114	<.001	0.131	0.17	0.287	0.017	0.775	0.158
			(0.004 - 0.007)	(0.005 - 0.03)	(-0.001 - 0.012)	(0.002 - 0.019)	(-0.001 - 0.01)	(0.003 - 0.007)	(-0.001 - 0.009)	(-0.012 - 0.058)	(-0.005 - 0.015)	(0.004 - 0.028)	(-0.01 - 0.008)	(-0.002 - 0.009)
	C3	P-value	0.767	0.79	0.781	0.757	0.782	0.762	0.663	0.673	0.67	0.614	0.687	0.659
		95% CI	<.001	0.053	0.1	<.01	0.113	<.001	0.429	0.417	0.863	<.01	0.865	0.64
			(0.005 - 0.008)	(-0.0 - 0.024)	(-0.001 - 0.009)	(0.007 - 0.024)	(-0.002 - 0.013)	(0.004 - 0.008)	(-0.002 - 0.004)	(-0.017 - 0.037)	(-0.015 - 0.018)	(0.01 - 0.042)	(-0.015 - 0.013)	(-0.003 - 0.005)
Overall SOFA score increase	R		0.761	0.747	0.756	0.762	0.762	0.763	0.644	0.624	0.664	0.635	0.653	0.642
	SpO2	P-value	0.76	0.74	0.752	0.758	0.763	0.762	0.641	0.577	0.653	0.619	0.652	0.642
		95% CI	0.055	0.353	0.174	0.413	0.715	0.187	0.081	<.01	0.178	0.228	0.915	0.962
			(-0.0 - 0.003)	(-0.009 - 0.024)	(-0.002 - 0.01)	(-0.006 - 0.013)	(-0.006 - 0.005)	(-0.001 - 0.003)	(-0.001 - 0.007)	(0.017 - 0.076)	(-0.006 - 0.029)	(-0.012 - 0.044)	(-0.012 - 0.013)	(-0.005 - 0.004)
	C1	P-value	0.76	0.75	0.751	0.753	0.761	0.762	0.642	0.61	0.664	0.62	0.649	0.641
		95% CI	0.048	0.584	0.036	0.013	0.778	0.29	0.102	0.353	0.989	0.166	0.513	0.546
			(0.0 - 0.003)	(-0.014 - 0.009)	(0.0 - 0.009)	(0.003 - 0.016)	(-0.005 - 0.007)	(-0.001 - 0.002)	(-0.001 - 0.005)	(-0.018 - 0.046)	(-0.016 - 0.017)	(-0.007 - 0.036)	(-0.009 - 0.016)	(-0.002 - 0.004)
	C2	P-value	0.758	0.747	0.753	0.758	0.762	0.759	0.638	0.613	0.671	0.623	0.656	0.634
		95% CI	<.001	0.964	0.342	0.303	0.919	<.001	<.01	0.504	0.367	0.324	0.732	<.001
			(0.002 - 0.005)	(-0.015 - 0.014)	(-0.003 - 0.008)	(-0.004 - 0.01)	(-0.004 - 0.005)	(0.003 - 0.005)	(0.003 - 0.009)	(-0.023 - 0.044)	(-0.024 - 0.01)	(-0.014 - 0.037)	(-0.022 - 0.016)	(0.005 - 0.011)
	C3	P-value	0.758	0.747	0.749	0.753	0.763	0.759	0.64	0.602	0.66	0.622	0.652	0.639
		95% CI	<.001	0.948	0.087	0.127	0.787	<.001	0.123	0.13	0.68	0.313	0.953	0.194
			(0.002 - 0.005)	(-0.014 - 0.013)	(-0.001 - 0.015)	(-0.003 - 0.02)	(-0.006 - 0.005)	(0.002 - 0.005)	(-0.001 - 0.008)	(-0.008 - 0.051)	(-0.018 - 0.026)	(-0.014 - 0.04)	(-0.011 - 0.012)	(-0.002 - 0.007)

Table B.10: F1-score and accuracy in the test set with XGBoost, in the Blood-gas and Oximetry study.

Task	Feature		F1-score						Accuracy					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.486	0.548	0.475	0.542	0.498	0.478	0.705	0.708	0.714	0.696	0.726	0.702
	SpO2		0.482	0.513	0.468	0.544	0.477	0.477	0.703	0.684	0.715	0.696	0.718	0.7
		P-value	0.046	0.022	0.215	0.809	<.01	0.683	0.168	<.01	0.631	0.981	0.073	0.357
		95% CI	(0.0 - 0.008)	(0.006 - 0.064)	(-0.005 - 0.02)	(-0.017 - 0.014)	(0.007 - 0.034)	(-0.005 - 0.008)	(-0.001 - 0.006)	(0.008 - 0.039)	(-0.007 - 0.004)	(-0.009 - 0.009)	(-0.001 - 0.018)	(-0.003 - 0.006)
	C1		0.484	0.545	0.475	0.545	0.489	0.476	0.705	0.712	0.714	0.698	0.722	0.702
		P-value	0.539	0.879	0.945	0.739	0.212	0.533	0.866	0.718	1	0.689	0.39	0.917
		95% CI	(-0.005 - 0.008)	(-0.038 - 0.043)	(-0.013 - 0.012)	(-0.02 - 0.015)	(-0.006 - 0.023)	(-0.004 - 0.008)	(-0.004 - 0.004)	(-0.029 - 0.021)	(-0.005 - 0.005)	(-0.012 - 0.008)	(-0.005 - 0.012)	(-0.004 - 0.004)
	C2		0.487	0.556	0.474	0.54	0.488	0.481	0.706	0.714	0.712	0.693	0.72	0.705
		P-value	0.646	0.519	0.836	0.794	0.07	0.361	0.613	0.292	0.622	0.465	0.09	0.226
		95% CI	(-0.006 - 0.004)	(-0.033 - 0.018)	(-0.012 - 0.015)	(-0.017 - 0.022)	(-0.001 - 0.021)	(-0.007 - 0.003)	(-0.004 - 0.003)	(-0.02 - 0.007)	(-0.007 - 0.01)	(-0.007 - 0.013)	(-0.001 - 0.012)	(-0.006 - 0.002)
	C3		0.487	0.566	0.478	0.541	0.485	0.48	0.706	0.718	0.718	0.699	0.721	0.704
		P-value	0.725	0.277	0.497	0.866	0.097	0.668	0.333	0.284	0.231	0.576	0.304	0.429
		95% CI	(-0.006 - 0.004)	(-0.054 - 0.017)	(-0.014 - 0.007)	(-0.019 - 0.022)	(-0.003 - 0.029)	(-0.008 - 0.005)	(-0.004 - 0.001)	(-0.032 - 0.011)	(-0.012 - 0.003)	(-0.017 - 0.01)	(-0.005 - 0.014)	(-0.005 - 0.002)
Respiratory SOFA score	R		0.661	0.672	0.656	0.66	0.676	0.657	0.716	0.754	0.736	0.716	0.728	0.711
	SpO2		0.657	0.655	0.648	0.654	0.674	0.654	0.713	0.745	0.734	0.714	0.726	0.707
		P-value	0.017	0.014	0.084	0.227	0.599	0.098	<.01	0.011	0.585	0.517	0.658	<.01
		95% CI	(0.001 - 0.007)	(0.004 - 0.031)	(-0.001 - 0.019)	(-0.004 - 0.016)	(-0.008 - 0.012)	(-0.001 - 0.007)	(0.001 - 0.006)	(0.003 - 0.016)	(-0.006 - 0.01)	(-0.005 - 0.009)	(-0.007 - 0.01)	(0.001 - 0.007)
	C1		0.657	0.657	0.651	0.646	0.671	0.654	0.712	0.745	0.728	0.705	0.725	0.707
		P-value	<.01	0.041	0.153	0.018	0.222	<.01	<.001	0.046	0.029	0.014	0.209	<.001
		95% CI	(0.002 - 0.006)	(0.001 - 0.031)	(-0.002 - 0.013)	(0.003 - 0.025)	(-0.003 - 0.013)	(0.001 - 0.004)	(0.003 - 0.007)	(0.0 - 0.017)	(0.001 - 0.015)	(0.003 - 0.018)	(-0.002 - 0.009)	(0.003 - 0.005)
	C2		0.655	0.654	0.646	0.647	0.676	0.651	0.71	0.746	0.726	0.706	0.727	0.706
		P-value	<.01	0.05	<.01	<.01	0.952	<.01	<.001	0.052	0.01	<.01	0.797	<.001
		95% CI	(0.003 - 0.009)	(0.0 - 0.038)	(0.003 - 0.018)	(0.005 - 0.021)	(-0.007 - 0.008)	(0.002 - 0.009)	(0.004 - 0.008)	(-0.0 - 0.016)	(0.003 - 0.018)	(0.003 - 0.016)	(-0.006 - 0.008)	(0.003 - 0.008)
	C3		0.656	0.663	0.648	0.64	0.675	0.652	0.71	0.75	0.728	0.702	0.726	0.705
		P-value	<.001	0.192	0.099	<.01	0.869	<.01	<.001	0.398	0.024	<.01	0.605	<.001
		95% CI	(0.003 - 0.008)	(-0.006 - 0.024)	(-0.002 - 0.019)	(0.008 - 0.032)	(-0.009 - 0.01)	(0.002 - 0.008)	(0.004 - 0.008)	(-0.005 - 0.012)	(0.001 - 0.015)	(0.005 - 0.022)	(-0.005 - 0.008)	(0.003 - 0.008)
Overall SOFA score increase	R		0.52	0.498	0.508	0.516	0.533	0.521	0.717	0.714	0.699	0.726	0.716	0.719
	SpO2		0.515	0.472	0.507	0.508	0.532	0.516	0.713	0.705	0.701	0.725	0.715	0.713
		P-value	<.001	0.025	0.744	0.376	0.857	<.01	<.01	0.257	0.42	0.835	0.958	<.001
		95% CI	(0.003 - 0.007)	(0.004 - 0.049)	(-0.01 - 0.013)	(-0.011 - 0.026)	(-0.009 - 0.011)	(0.003 - 0.008)	(0.002 - 0.006)	(-0.009 - 0.028)	(-0.01 - 0.005)	(-0.01 - 0.012)	(-0.008 - 0.009)	(0.003 - 0.008)
	C1		0.517	0.489	0.506	0.502	0.531	0.519	0.714	0.709	0.695	0.717	0.714	0.717
		P-value	0.034	0.287	0.527	0.039	0.456	0.158	0.032	0.496	0.205	0.06	0.718	0.152
		95% CI	(0.0 - 0.006)	(-0.009 - 0.027)	(-0.006 - 0.012)	(0.001 - 0.026)	(-0.005 - 0.01)	(-0.001 - 0.005)	(0.0 - 0.005)	(-0.011 - 0.021)	(-0.002 - 0.009)	(-0.0 - 0.017)	(-0.006 - 0.008)	(-0.001 - 0.004)
	C2		0.514	0.485	0.511	0.507	0.532	0.514	0.713	0.702	0.698	0.721	0.714	0.715
		P-value	<.001	0.098	0.69	0.055	0.863	<.001	<.001	0.128	0.81	0.052	0.724	<.001
		95% CI	(0.004 - 0.007)	(-0.003 - 0.03)	(-0.016 - 0.011)	(-0.0 - 0.018)	(-0.009 - 0.011)	(0.005 - 0.008)	(0.002 - 0.005)	(-0.004 - 0.029)	(-0.008 - 0.01)	(-0.0 - 0.01)	(-0.008 - 0.012)	(0.002 - 0.005)
	C3		0.514	0.49	0.502	0.503	0.533	0.516	0.712	0.713	0.693	0.717	0.716	0.714
		P-value	<.01	0.075	0.335	0.157	0.864	<.01	<.01	0.741	0.187	0.126	0.984	<.001
		95% CI	(0.002 - 0.009)	(-0.001 - 0.018)	(-0.008 - 0.02)	(-0.006 - 0.032)	(-0.008 - 0.01)	(0.002 - 0.008)	(0.002 - 0.007)	(-0.008 - 0.011)	(-0.003 - 0.015)	(-0.003 - 0.021)	(-0.011 - 0.011)	(0.002 - 0.006)

Table B.11: Performance metrics in the test set per disparity groups with LR, in the Blood-gas and Oximetry study.

Task	Feature		AUROC				Recall				F1-score				Accuracy			
			(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)
In-hospital mortality	R		0.721	0.7	0.701	0.718	0.693	0.609	0.592	0.669	0.573	0.464	0.45	0.502	0.652	0.658	0.68	0.655
	SpO2	P-value	0.722	0.701	0.701	0.716	0.755	0.638	0.592	0.638	0.572	0.467	0.448	0.498	0.621	0.646	0.676	0.665
		95% CI	0.804	0.048	0.859	0.064	<.001	<.001	1	<.001	0.877	0.11	0.02	0.088	<.001	<.001	0.058	<.001
			(-0.003 - 0.002)	(-0.002 - -0.0)	(-0.001 - 0.001)	(-0.0 - 0.004)	(-0.074 - -0.048)	(-0.035 - -0.022)	(-0.007 - -0.007)	(0.022 - 0.04)	(-0.008 - 0.009)	(-0.007 - 0.001)	(0.0 - 0.005)	(0.001 - 0.01)	(0.021 - 0.042)	(0.008 - 0.016)	(-0.0 - 0.007)	(-0.013 - -0.006)
	C1	P-value	0.703	0.699	0.708	0.699	0.697	0.619	0.597	0.642	0.548	0.469	0.443	0.522	0.63	0.655	0.684	0.652
		95% CI	0.26	0.871	0.011	0.081	0.843	0.418	0.521	0.199	0.113	0.563	0.094	0.12	0.097	0.57	0.031	0.563
			(-0.016 - 0.054)	(-0.011 - 0.013)	(0.012 - 0.002)	(-0.003 - 0.04)	(-0.047 - 0.039)	(-0.039 - 0.018)	(-0.018 - 0.01)	(-0.017 - 0.071)	(-0.007 - 0.058)	(-0.022 - 0.013)	(-0.001 - 0.015)	(-0.045 - 0.006)	(-0.005 - 0.049)	(-0.007 - 0.013)	(-0.008 - -0.0)	(-0.01 - 0.017)
	C2	P-value	0.705	0.707	0.703	0.703	0.711	0.605	0.596	0.658	0.548	0.45	0.454	0.535	0.627	0.676	0.675	0.65
		95% CI	0.281	0.068	0.55	0.18	0.278	0.597	0.635	0.484	0.077	0.062	0.391	<.01	0.05	<.01	0.143	0.422
			(-0.016 - 0.048)	(-0.014 - 0.001)	(-0.008 - 0.004)	(-0.008 - 0.036)	(-0.054 - 0.017)	(-0.013 - 0.022)	(-0.019 - 0.012)	(-0.023 - 0.046)	(-0.003 - 0.054)	(-0.001 - 0.03)	(-0.014 - 0.006)	(-0.054 - 0.012)	(0.0 - 0.05)	(-0.028 - 0.007)	(-0.002 - 0.012)	(-0.009 - 0.021)
	C3	P-value	0.712	0.701	0.708	0.714	0.71	0.591	0.606	0.685	0.55	0.446	0.459	0.559	0.632	0.674	0.679	0.651
		95% CI	0.491	0.935	0.072	0.625	0.418	0.046	0.093	0.147	0.116	0.044	0.096	<.001	0.101	0.01	0.628	0.402
			(-0.02 - 0.039)	(-0.011 - 0.01)	(-0.015 - 0.001)	(-0.014 - 0.021)	(-0.06 - 0.027)	(0.0 - 0.036)	(-0.031 - 0.003)	(-0.039 - 0.007)	(-0.007 - 0.053)	(0.001 - 0.036)	(-0.021 - 0.002)	(-0.072 - 0.041)	(-0.005 - 0.046)	(-0.027 - 0.005)	(-0.005 - 0.007)	(-0.007 - 0.017)
Respiratory SOFA score	R		0.757	0.756	0.752	0.754	0.693	0.704	0.654	0.749	0.69	0.67	0.624	0.703	0.702	0.703	0.706	0.69
	SpO2	P-value	0.75	0.755	0.751	0.755	0.783	0.726	0.657	0.698	0.7	0.672	0.625	0.693	0.678	0.698	0.705	0.698
		95% CI	0.074	0.44	0.07	0.287	<.001	<.001	0.13	<.001	<.01	0.052	0.37	<.001	<.001	<.001	0.025	<.001
			(-0.001 - 0.014)	(-0.001 - 0.002)	(-0.0 - 0.003)	(-0.003 - 0.001)	(-0.105 - 0.076)	(-0.028 - 0.017)	(-0.006 - 0.001)	(-0.046 - 0.055)	(-0.017 - 0.004)	(-0.006 - 0.0)	(-0.002 - 0.001)	(-0.007 - 0.012)	(0.013 - 0.034)	(0.003 - 0.007)	(0.0 - 0.001)	(-0.011 - 0.004)
	C1	P-value	0.743	0.752	0.748	0.764	0.776	0.708	0.656	0.695	0.714	0.666	0.619	0.715	0.681	0.696	0.704	0.704
		95% CI	<.001	0.291	<.01	0.014	<.001	0.376	0.631	<.001	0.012	0.519	0.048	0.092	<.01	0.033	0.058	<.01
			(0.008 - 0.02)	(-0.004 - 0.011)	(0.001 - 0.007)	(-0.017 - 0.003)	(-0.095 - 0.071)	(-0.015 - 0.006)	(-0.009 - 0.006)	(-0.039 - 0.069)	(-0.04 - 0.006)	(-0.007 - 0.014)	(0.0 - 0.01)	(-0.026 - 0.002)	(0.01 - 0.032)	(0.001 - 0.014)	(-0.0 - 0.004)	(-0.022 - 0.006)
	C2	P-value	0.744	0.751	0.751	0.765	0.776	0.675	0.666	0.722	0.712	0.629	0.64	0.742	0.681	0.701	0.702	0.71
		95% CI	<.01	0.319	0.447	0.024	<.001	<.001	0.017	<.01	0.012	<.001	<.001	<.001	<.01	0.564	0.024	<.01
			(0.006 - 0.02)	(-0.005 - 0.014)	(-0.003 - 0.006)	(-0.02 - 0.002)	(-0.095 - 0.071)	(-0.018 - 0.039)	(-0.021 - 0.003)	(-0.012 - 0.042)	(-0.038 - 0.006)	(-0.029 - 0.051)	(-0.022 - -0.01)	(-0.052 - 0.026)	(0.01 - 0.032)	(-0.006 - 0.011)	(-0.001 - 0.007)	(-0.03 - 0.011)
	C3	P-value	0.734	0.752	0.746	0.757	0.778	0.676	0.666	0.724	0.696	0.62	0.66	0.747	0.67	0.707	0.697	0.707
		95% CI	<.001	0.496	0.05	0.574	<.001	<.001	0.032	<.01	0.356	<.001	<.001	<.001	<.001	0.302	<.01	<.001
			(0.015 - 0.03)	(-0.007 - 0.013)	(-0.0 - 0.013)	(-0.016 - 0.009)	(-0.095 - 0.075)	(-0.017 - 0.038)	(-0.023 - 0.001)	(-0.012 - 0.037)	(-0.021 - 0.008)	(-0.038 - 0.06)	(-0.043 - 0.029)	(-0.056 - 0.034)	(-0.021 - 0.044)	(-0.011 - 0.004)	(-0.003 - 0.014)	(-0.025 - 0.009)
Overall SOFA score increase	R		0.722	0.732	0.732	0.731	0.646	0.668	0.676	0.697	0.53	0.487	0.484	0.501	0.666	0.674	0.669	0.65
	SpO2	P-value	0.724	0.732	0.733	0.728	0.736	0.698	0.679	0.648	0.533	0.484	0.483	0.499	0.623	0.656	0.666	0.672
		95% CI	0.104	0.84	0.193	0.013	<.001	<.001	0.064	<.001	0.536	0.076	0.322	0.093	<.001	<.001	<.01	<.001
			(-0.004 - 0.0)	(-0.001 - 0.001)	(-0.001 - 0.0)	(0.001 - 0.005)	(-0.112 - 0.069)	(-0.035 - 0.026)	(-0.006 - 0.0)	(-0.044 - 0.055)	(-0.013 - 0.007)	(-0.0 - 0.005)	(-0.001 - 0.002)	(-0.0 - 0.004)	(0.035 - 0.051)	(0.014 - 0.022)	(0.001 - 0.004)	(-0.025 - -0.02)
	C1	P-value	0.732	0.73	0.736	0.727	0.678	0.659	0.708	0.59	0.51	0.479	0.484	0.513	0.663	0.675	0.651	0.696
		95% CI	0.195	0.629	0.297	0.518	0.074	0.436	<.001	<.001	0.107	0.106	0.811	0.141	0.548	0.802	<.001	<.001
			(-0.024 - 0.006)	(-0.008 - 0.013)	(-0.01 - 0.003)	(-0.008 - 0.016)	(-0.068 - 0.004)	(-0.016 - 0.034)	(-0.043 - -0.02)	(-0.084 - 0.131)	(-0.005 - 0.045)	(-0.002 - 0.017)	(-0.009 - 0.007)	(-0.028 - 0.005)	(-0.009 - 0.015)	(-0.009 - 0.007)	(-0.012 - 0.024)	(-0.056 - 0.036)
	C2	P-value	0.73	0.735	0.733	0.726	0.672	0.699	0.681	0.592	0.505	0.485	0.484	0.521	0.66	0.659	0.665	0.692
		95% CI	0.289	0.639	0.946	0.381	0.125	0.014	0.336	<.001	0.052	0.82	0.879	0.026	0.278	<.01	0.19	<.001
			(-0.023 - 0.008)	(-0.012 - 0.008)	(-0.007 - 0.006)	(-0.007 - 0.016)	(-0.062 - 0.009)	(-0.055 - 0.008)	(-0.015 - 0.006)	(-0.079 - 0.132)	(-0.0 - 0.051)	(-0.009 - 0.012)	(-0.011 - 0.009)	(-0.037 - 0.003)	(-0.005 - 0.016)	(-0.005 - 0.025)	(-0.002 - 0.01)	(-0.053 - 0.031)
	C3	P-value	0.735	0.736	0.727	0.723	0.67	0.69	0.68	0.623	0.511	0.474	0.49	0.546	0.67	0.663	0.658	0.677
		95% CI	0.033	0.393	0.237	0.225	0.184	0.032	0.501	<.001	0.108	0.026	0.23	<.001	0.452	<.01	0.013	<.001
			(-0.024 - -0.001)	(-0.012 - 0.005)	(-0.004 - 0.014)	(-0.006 - 0.021)	(-0.062 - 0.014)	(-0.043 - 0.003)	(-0.017 - 0.009)	(0.05 - 0.098)	(-0.005 - 0.044)	(0.002 - 0.024)	(-0.017 - 0.005)	(-0.057 - 0.033)	(-0.016 - 0.008)	(0.004 - 0.018)	(0.003 - 0.019)	(-0.039 - 0.015)

Table B.12: Performance metrics in the test set per disparity groups with RF, in the Blood-gas and Oximetry study.

Task	Feature		AUROC				Recall				F1-score				Accuracy			
			(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)
In-hospital mortality	R		0.741	0.728	0.727	0.746	0.144	0.108	0.098	0.141	0.239	0.184	0.168	0.228	0.69	0.769	0.786	0.754
	SpO2		0.741	0.726	0.728	0.748	0.188	0.126	0.1	0.128	0.292	0.209	0.172	0.213	0.694	0.77	0.787	0.755
		P-value	0.891	0.404	0.456	0.465	<.001	<.01	0.101	<.01	<.01	0.013	0.099	<.01	0.356	0.405	0.381	0.177
		95% CI	(-0.006 - 0.007)	(-0.003 - 0.007)	(-0.004 - 0.002)	(-0.007 - 0.003)	(-0.065 - - 0.023)	(-0.03 - - 0.006)	(-0.006 - 0.001)	(0.006 - 0.019)	(- - - 0.025)	(-0.081 - - 0.007)	(-0.044 - - 0.001)	(-0.01 - 0.024)	(0.006 - 0.005)	(-0.012 - 0.002)	(-0.004 - 0.001)	(-0.002 - 0.001)
	C1		0.724	0.718	0.741	0.718	0.18	0.111	0.097	0.12	0.283	0.187	0.167	0.201	0.707	0.764	0.797	0.717
		P-value	0.114	0.052	<.001	0.016	<.01	0.71	0.772	0.019	0.01	0.819	0.926	0.051	0.181	0.481	<.001	<.01
		95% CI	(-0.005 - 0.04)	(-0.0 - 0.021)	(-0.018 - -0.01)	(0.007 - 0.049)	(- - 0.014)	(-0.058 - - 0.015)	(-0.021 - 0.005)	(0.004 - 0.036)	(-0.075 - - 0.013)	(-0.03 - 0.024)	(-0.007 - 0.007)	(-0.0 - 0.055)	(-0.042 - 0.009)	(-0.01 - 0.02)	(-0.015 - - 0.007)	(0.018 - 0.055)
	C2		0.722	0.735	0.732	0.721	0.18	0.106	0.102	0.118	0.283	0.18	0.175	0.196	0.711	0.79	0.782	0.704
		P-value	0.078	0.118	0.083	0.029	<.01	0.775	0.219	0.016	0.016	0.739	0.185	0.032	0.117	<.01	0.239	<.001
		95% CI	(-0.003 - 0.041)	(-0.016 - 0.002)	(-0.011 - 0.001)	(0.003 - 0.046)	(- - 0.011)	(-0.059 - - 0.017)	(-0.013 - 0.003)	(0.005 - 0.04)	(-0.077 - -0.01)	(-0.02 - 0.027)	(-0.018 - 0.004)	(0.003 - 0.06)	(-0.048 - 0.006)	(0.033 - - 0.009)	(0.003 - 0.01)	(0.031 - 0.068)
	C3		0.725	0.728	0.739	0.735	0.198	0.104	0.101	0.156	0.303	0.177	0.173	0.25	0.712	0.787	0.782	0.7
		P-value	0.177	1	0.013	0.126	<.01	0.463	0.252	0.143	<.01	0.443	0.269	0.142	0.071	<.01	0.183	<.001
		95% CI	(-0.009 - 0.04)	(-0.013 - 0.013)	(-0.02 - - 0.003)	(-0.004 - 0.025)	(-0.082 - - 0.025)	(-0.007 - - 0.015)	(-0.01 - 0.003)	(-0.036 - 0.006)	(-0.1 - - 0.028)	(-0.012 - 0.025)	(-0.015 - 0.005)	(-0.055 - 0.009)	(-0.046 - 0.002)	(-0.027 - - 0.008)	(-0.002 - 0.01)	(0.034 - 0.074)
Respiratory SOFA score	R		0.771	0.771	0.771	0.77	0.543	0.53	0.505	0.657	0.64	0.614	0.587	0.69	0.706	0.716	0.734	0.711
	SpO2		0.757	0.766	0.768	0.77	0.687	0.627	0.513	0.541	0.689	0.654	0.588	0.634	0.702	0.716	0.732	0.695
		P-value	<.01	<.001	<.001	1	<.001	<.001	<.001	<.001	<.001	<.001	0.257	<.001	0.493	0.766	0.016	<.001
		95% CI	(0.007 - 0.02)	(0.003 - 0.007)	(0.003 - 0.005)	(-0.002 - 0.002)	(- - 0.123)	(- - 0.087)	(- - 0.004)	(0.109 - 0.124)	(- - 0.031)	(- - 0.031)	(- 0.001)	(0.048 - 0.063)	(-0.01 - 0.02)	(- 0.004)	(0.001 - 0.004)	(0.011 - 0.021)
	C1		0.747	0.765	0.767	0.772	0.696	0.582	0.499	0.543	0.699	0.633	0.578	0.647	0.694	0.711	0.733	0.683
		P-value	<.001	0.099	0.031	0.608	<.001	<.001	0.035	<.001	<.001	<.01	0.011	<.001	0.03	0.061	0.411	<.001
		95% CI	(0.017 - 0.031)	(- 0.001 - 0.012)	(0.001 - 0.009)	(- 0.008 - 0.005)	(- - 0.135)	(- - 0.037)	(0.0 - 0.011)	(0.106 - 0.122)	(- 0.081 - - 0.039)	(- 0.031 - - 0.008)	(0.003 - 0.015)	(0.036 - 0.051)	(0.002 - 0.024)	(-0.0 - 0.01)	(- 0.003 - 0.007)	(0.02 - 0.036)
	C2		0.748	0.768	0.766	0.771	0.687	0.543	0.508	0.555	0.695	0.602	0.591	0.661	0.692	0.73	0.721	0.672
		P-value	<.001	0.398	0.055	0.882	<.001	0.032	0.653	<.001	<.001	0.017	0.379	<.001	0.034	<.01	<.001	<.001
		95% CI	(0.014 - 0.033)	(- 0.004 - 0.009)	(-0.0 - 0.011)	(- 0.009 - 0.008)	(- - 0.126)	(- - 0.001)	(- 0.014 - 0.009)	(0.086 - 0.117)	(- 0.076 - - 0.034)	(0.003 - 0.022)	(- 0.014 - 0.006)	(0.016 - 0.041)	(0.001 - 0.027)	(- 0.021 - - 0.007)	(0.009 - 0.019)	(0.023 - 0.054)
	C3		0.739	0.769	0.76	0.765	0.696	0.524	0.527	0.598	0.683	0.587	0.612	0.692	0.685	0.74	0.705	0.683
		P-value	<.001	0.663	<.01	0.31	<.001	0.367	<.001	<.001	<.001	<.01	<.001	0.697	<.01	<.001	<.001	<.001
		95% CI	(0.023 - 0.04)	(- 0.007 - 0.011)	(0.005 - 0.017)	(- 0.006 - 0.016)	(- - 0.138)	(- 0.009 - 0.023)	(- - 0.012)	(0.043 - 0.076)	(- 0.062 - - 0.024)	(0.011 - 0.043)	(- 0.034 - - 0.015)	(- 0.018 - 0.012)	(0.009 - 0.033)	(- 0.033 - - 0.015)	(0.023 - 0.036)	(0.017 - 0.038)
Overall SOFA score increase	R		0.738	0.747	0.756	0.751	0.17	0.151	0.18	0.184	0.275	0.247	0.285	0.29	0.74	0.787	0.793	0.773
	SpO2		0.729	0.747	0.754	0.75	0.213	0.157	0.177	0.165	0.327	0.254	0.281	0.268	0.744	0.787	0.792	0.773
		P-value	0.083	0.95	0.212	0.708	<.001	0.131	0.241	<.001	<.001	0.152	0.197	<.001	0.026	1	0.171	0.825
		95% CI	(-0.001 - 0.019)	(- 0.003 - 0.004)	(- 0.001 - 0.005)	(- 0.002 - 0.003)	(- - 0.028)	(- 0.014 - 0.002)	(- 0.002 - 0.007)	(-0.013 - 0.027)	(- 0.071 - - 0.032)	(- 0.017 - 0.003)	(- 0.002 - 0.01)	(-0.013 - 0.032)	(- 0.007 - - 0.001)	(- 0.002 - 0.002)	(-0.0 - 0.002)	(- 0.002 - 0.002)
	C1		0.744	0.749	0.757	0.746	0.159	0.141	0.202	0.129	0.26	0.234	0.311	0.222	0.765	0.791	0.792	0.754
		P-value	0.488	0.612	0.733	0.26	0.402	0.172	<.001	<.001	0.375	0.23	<.001	<.001	<.001	0.229	0.334	<.01
		95% CI	(-0.021 - 0.011)	(- 0.012 - 0.007)	(- 0.006 - 0.004)	(- 0.004 - 0.014)	(- 0.016 - 0.037)	(- 0.005 - 0.026)	(- 0.031 - - 0.014)	(- - 0.066)	(- 0.021 - 0.051)	(-0.01 - 0.036)	(- 0.036 - - 0.015)	(- 0.053 - 0.084)	(- 0.037 - - 0.014)	(-0.01 - 0.003)	(- 0.002 - 0.005)	(0.007 - 0.031)
	C2		0.74	0.754	0.752	0.746	0.162	0.185	0.176	0.128	0.263	0.29	0.279	0.221	0.767	0.791	0.79	0.743
		P-value	0.823	0.108	0.093	0.255	0.464	<.001	0.405	<.001	0.437	<.01	0.368	<.001	<.001	0.182	0.163	<.001
		95% CI	(-0.022 - 0.018)	(- 0.016 - 0.002)	(- 0.001 - 0.009)	(- 0.004 - 0.013)	(- 0.016 - 0.032)	(- 0.048 - 0.019)	(- 0.006 - 0.014)	(-0.046 - 0.067)	(- 0.021 - 0.045)	(- 0.063 - - 0.022)	(- 0.008 - 0.02)	(-0.054 - 0.086)	(- 0.038 - - 0.016)	(- 0.011 - 0.002)	(- 0.002 - 0.008)	(0.017 - 0.042)
	C3		0.742	0.756	0.746	0.74	0.173	0.179	0.172	0.148	0.277	0.283	0.274	0.248	0.768	0.8	0.78	0.719
		P-value	0.517	0.067	0.038	0.033	0.696	<.01	0.102	<.001	0.857	<.01	0.105	<.01	<.001	<.001	<.001	<.001
		95% CI	(-0.018 - 0.01)	(- 0.018 - 0.001)	(0.001 - 0.018)	(0.001 - 0.021)	(- 0.023 - 0.016)	(- 0.044 - - 0.013)	(- 0.002 - 0.018)	(0.02 - 0.052)	(-0.03 - 0.026)	(- 0.058 - - 0.015)	(- 0.003 - 0.025)	(0.02 - 0.066)	(-0.04 - - 0.017)	(- 0.019 - - 0.008)	(0.007 - 0.019)	(0.037 - 0.07)

Table B.13: Performance metrics in the test set per disparity groups with XGBoost, in the Blood-gas and Oximetry study.

Task	Feature		AUROC				Recall				F1-score				Accuracy			
			(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)	(-∞, -3)	[-3, 0)	[0, 3)	[3, +∞)
In-hospital mortality	R		0.731	0.718	0.723	0.738	0.649	0.569	0.553	0.627	0.573	0.479	0.461	0.516	0.674	0.699	0.715	0.694
	SpO2		0.731	0.718	0.722	0.731	0.714	0.584	0.551	0.587	0.577	0.478	0.459	0.505	0.648	0.69	0.712	0.701
		P-value	0.913	0.9	0.61	0.023	<.001	0.027	0.8	<.001	0.517	0.931	0.329	0.011	<.001	0.018	0.123	<.01
		95% CI	(-0.008	(-0.007	(-0.003	(0.001 - 0.012)	(-0.085	(-0.028	(-0.008	(0.024 - 0.056)	(-0.017 - 0.009)	(-0.01 - 0.011)	(-0.003 - 0.008)	(0.003 - 0.019)	(0.017 - 0.037)	(0.002 - 0.015)	(-0.001 - 0.006)	(-0.011 - 0.002)
	C1		0.723	0.714	0.735	0.699	0.679	0.57	0.553	0.593	0.559	0.475	0.461	0.513	0.656	0.692	0.728	0.668
		P-value	0.359	0.332	0.018	0.014	0.103	0.946	0.985	0.09	0.39	0.672	0.889	0.796	0.052	0.084	<.001	<.01
		95% CI	(-0.011 - 0.027)	(-0.004 - 0.012)	(-0.02 - - 0.002)	(0.01 - 0.066)	(-0.067 - 0.007)	(-0.017 - 0.016)	(-0.012 - 0.012)	(-0.007 - 0.076)	(-0.021 - 0.049)	(-0.014 - 0.021)	(-0.007 - 0.008)	(-0.022 - 0.027)	(-0.0 - 0.038)	(0.001 - 0.015)	(-0.017 - - 0.008)	(0.01 - 0.042)
	C2		0.723	0.73	0.728	0.706	0.669	0.563	0.563	0.607	0.554	0.467	0.47	0.534	0.659	0.718	0.713	0.675
		P-value	0.287	<.01	0.092	0.033	0.107	0.446	0.04	0.238	0.207	0.148	<.01	0.229	0.082	<.001	0.502	0.061
		95% CI	(-0.008 - 0.024)	(-0.02 - - 0.004)	(-0.01 - 0.001)	(0.003 - 0.06)	(-0.044 - 0.005)	(-0.011 - 0.023)	(-0.02 - - 0.001)	(-0.016 - 0.056)	(-0.012 - 0.049)	(-0.005 - 0.029)	(-0.015 - 0.003)	(-0.048 - 0.013)	(-0.002 - 0.033)	(-0.028 - 0.011)	(-0.005 - 0.009)	(-0.001 - 0.04)
	C3		0.723	0.723	0.731	0.719	0.677	0.556	0.564	0.632	0.559	0.463	0.472	0.554	0.661	0.713	0.716	0.672
		P-value	0.406	0.124	0.089	0.014	0.109	0.158	0.149	0.702	0.358	0.102	0.063	<.001	0.156	<.001	0.701	<.001
		95% CI	(-0.013 - 0.03)	(-0.012 - 0.002)	(-0.018 - 0.002)	(0.005 - 0.032)	(-0.063 - 0.008)	(-0.006 - 0.032)	(-0.029 - 0.005)	(-0.029 - 0.02)	(-0.019 - 0.047)	(-0.004 - 0.035)	(-0.023 - 0.001)	(-0.051 - 0.025)	(-0.006 - 0.033)	(-0.021 - 0.008)	(-0.005 - 0.004)	(0.012 - 0.032)
Respiratory SOFA score	R		0.768	0.768	0.771	0.769	0.641	0.641	0.619	0.759	0.676	0.657	0.628	0.71	0.706	0.714	0.727	0.697
	SpO2		0.764	0.767	0.769	0.768	0.796	0.722	0.625	0.66	0.699	0.674	0.628	0.686	0.671	0.702	0.723	0.705
		P-value	0.451	0.566	0.011	0.375	<.001	<.001	<.001	<.001	<.01	<.001	0.534	<.001	<.001	<.01	<.001	<.01
		95% CI	(-0.007 - 0.014)	(-0.003 - 0.005)	(0.001 - 0.004)	(-0.002 - 0.005)	(-0.177 - 0.134)	(-0.088 - 0.075)	(-0.009 - 0.004)	(-0.089 - 0.107)	(-0.038 - 0.008)	(-0.024 - -0.01)	(-0.001 - 0.002)	(0.018 - 0.03)	(0.021 - 0.048)	(0.006 - 0.019)	(0.002 - 0.005)	(-0.013 - - 0.003)
	C1		0.747	0.765	0.767	0.765	0.788	0.686	0.621	0.67	0.712	0.666	0.623	0.704	0.673	0.705	0.724	0.7
		P-value	<.001	0.31	<.01	0.31	<.001	<.001	0.383	<.001	<.001	0.059	<.01	0.335	<.001	<.001	0.01	0.442
		95% CI	(0.016 - 0.026)	(-0.003 - 0.01)	(0.001 - 0.007)	(-0.004 - 0.012)	(-0.162 - 0.133)	(-0.059 - 0.031)	(-0.007 - 0.003)	(-0.078 - 0.1)	(-0.052 - 0.019)	(-0.018 - 0.0)	(-0.003 - 0.009)	(-0.007 - 0.018)	(-0.022 - 0.044)	(-0.006 - 0.014)	(-0.001 - 0.005)	(-0.013 - 0.006)
	C2		0.748	0.768	0.766	0.769	0.783	0.65	0.626	0.676	0.709	0.633	0.636	0.723	0.674	0.717	0.715	0.702
		P-value	<.001	0.902	0.054	0.941	<.001	0.088	0.141	<.001	<.01	<.001	0.042	0.019	<.001	0.313	<.001	0.328
		95% CI	(0.014 - 0.025)	(-0.006 - 0.005)	(-0.0 - 0.01)	(-0.009 - 0.009)	(-0.161 - 0.124)	(-0.019 - 0.002)	(-0.018 - 0.003)	(-0.097)	(-0.053 - 0.014)	(-0.03)	(-0.014 - -0.0)	(-0.024 - 0.003)	(0.02 - 0.044)	(-0.008 - 0.003)	(-0.009 - 0.015)	(-0.016 - 0.006)
	C3		0.739	0.77	0.758	0.762	0.79	0.638	0.643	0.709	0.693	0.622	0.657	0.738	0.659	0.725	0.704	0.698
		P-value	<.001	0.552	<.001	0.189	<.001	0.407	<.001	<.001	0.045	<.001	<.001	<.001	<.001	<.01	<.001	0.737
		95% CI	(0.022 - 0.036)	(-0.01 - 0.006)	(0.007 - 0.019)	(-0.004 - 0.019)	(-0.165 - 0.135)	(-0.004 - 0.009)	(-0.033 - 0.015)	(-0.036 - 0.062)	(-0.034 - -0.0)	(0.026 - 0.044)	(-0.036 - 0.021)	(-0.04 - 0.015)	(0.036 - 0.056)	(-0.017 - 0.005)	(-0.017 - 0.029)	(-0.011 - 0.008)
Overall SOFA score increase	R		0.744	0.758	0.764	0.76	0.628	0.604	0.636	0.689	0.552	0.508	0.515	0.528	0.702	0.73	0.726	0.69
	SpO2		0.738	0.757	0.764	0.757	0.74	0.653	0.637	0.619	0.537	0.503	0.513	0.521	0.627	0.702	0.723	0.713
		P-value	0.153	0.341	1	0.117	<.001	<.001	0.704	<.001	0.024	0.07	0.256	0.021	<.001	<.001	0.136	<.001
		95% CI	(-0.015)	(-0.005)	(-0.002)	(-0.007)	(-0.092)	(-0.04)	(-0.004)	(0.06 - 0.079)	(0.002 - 0.027)	(-0.011)	(-0.007)	(0.001 - 0.014)	(0.068 - 0.082)	(0.024 - 0.032)	(-0.001 - 0.007)	(-0.028 - - 0.018)
	C1		0.751	0.756	0.767	0.752	0.697	0.61	0.659	0.592	0.528	0.503	0.517	0.533	0.676	0.726	0.714	0.718
		P-value	0.203	0.585	0.352	0.069	<.01	0.317	<.001	<.001	0.042	0.115	0.662	0.454	<.01	0.305	<.001	<.001
		95% CI	(-0.005)	(-0.011)	(-0.004)	(-0.018)	(-0.036)	(-0.02 - 0.007)	(-0.03 - 0.016)	(0.085 - 0.109)	(0.001 - 0.047)	(-0.002 - 0.012)	(-0.009 - 0.006)	(-0.018 - 0.009)	(0.012 - 0.039)	(-0.003 - 0.009)	(-0.007 - 0.016)	(-0.039 - - 0.018)
	C2		0.748	0.763	0.759	0.749	0.692	0.645	0.633	0.574	0.523	0.514	0.508	0.534	0.675	0.719	0.717	0.716
		P-value	0.464	0.107	0.055	0.036	<.01	<.001	0.438	<.001	0.026	0.079	0.041	0.374	<.001	0.01	<.01	<.001
		95% CI	(-0.007)	(-0.001)	(-0.0 - 0.01)	(0.001 - 0.021)	(-0.103 - 0.027)	(-0.054 - 0.028)	(-0.005 - 0.011)	(0.098 - 0.132)	(0.004 - 0.053)	(-0.011 - 0.001)	(0.0 - 0.015)	(-0.02 - 0.008)	(0.018 - 0.036)	(0.003 - 0.018)	(0.004 - 0.015)	(-0.035 - - 0.018)
	C3		0.753	0.764	0.753	0.748	0.701	0.642	0.631	0.622	0.528	0.506	0.509	0.559	0.678	0.725	0.707	0.694
		P-value	0.129	0.077	<.01	0.08	<.01	<.001	0.255	<.001	0.066	0.589	0.202	<.01	<.01	0.061	<.001	0.47
		95% CI	(-0.02 - 0.003)	(-0.012 - 0.001)	(0.004 - 0.018)	(-0.002 - 0.026)	(-0.113 - 0.034)	(-0.052 - 0.023)	(-0.004 - 0.015)	(-0.048 - 0.086)	(-0.002 - 0.05)	(-0.008 - 0.013)	(-0.004 - 0.016)	(-0.049 - 0.013)	(0.009 - 0.039)	(-0.0 - 0.01)	(-0.014 - 0.024)	(-0.018 - 0.009)

Table B.14: Performance metrics per No/Hidden Hypoxemia groups, in the test set, with LR.

Task	Feature		AUROC		Recall		F1-score		Accuracy	
			No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia
In-hospital mortality	R		0.704	0.726	0.608	0.777	0.462	0.58	0.671	0.627
	SpO2	P-value	0.704	0.729	0.609	0.726	0.46	0.58	0.668	0.652
		P-value	0.343	0.093	0.638	<.001	0.019	0.924	0.051	<.001
		95% CI	(-0.001 - 0.002)	(-0.008 - 0.001)	(-0.008 - 0.005)	(0.038 - 0.064)	(0.0 - 0.004)	(-0.01 - 0.009)	(-0.0 - 0.006)	(-0.034 - -0.016)
	C1	P-value	0.703	0.705	0.605	0.699	0.458	0.572	0.67	0.642
		P-value	0.086	0.119	0.362	<.001	<.01	0.623	0.595	0.182
		95% CI	(-0.0 - 0.003)	(-0.006 - 0.047)	(-0.003 - 0.007)	(0.051 - 0.104)	(0.001 - 0.006)	(-0.029 - 0.046)	(-0.002 - 0.003)	(-0.04 - 0.009)
	C2	P-value	0.703	0.704	0.602	0.703	0.457	0.57	0.671	0.64
		P-value	0.1	0.072	0.114	<.001	<.001	0.547	0.819	0.235
		95% CI	(-0.0 - 0.004)	(-0.002 - 0.046)	(-0.002 - 0.013)	(0.047 - 0.1)	(0.003 - 0.007)	(-0.025 - 0.045)	(-0.004 - 0.003)	(-0.035 - 0.01)
	C3	P-value	0.706	0.717	0.608	0.726	0.463	0.591	0.672	0.647
		P-value	0.012	0.085	0.734	<.001	0.475	0.1	0.414	<.01
		95% CI	(-0.003 - -0.0)	(-0.001 - 0.019)	(-0.006 - 0.004)	(0.03 - 0.071)	(-0.004 - 0.002)	(-0.025 - 0.003)	(-0.003 - 0.001)	(-0.033 - -0.007)
Respiratory SOFA score	R		0.754	0.729	0.679	0.852	0.649	0.753	0.702	0.662
	SpO2	P-value	0.75	0.765	0.679	0.729	0.647	0.753	0.701	0.712
		P-value	<.001	<.001	0.78	<.001	0.039	0.985	0.041	<.001
		95% CI	(0.003 - 0.005)	(-0.05 - -0.023)	(-0.003 - 0.004)	(0.098 - 0.147)	(0.0 - 0.003)	(-0.012 - 0.011)	(0.0 - 0.002)	(-0.061 - -0.037)
	C1	P-value	0.749	0.754	0.676	0.717	0.644	0.758	0.701	0.706
		P-value	<.001	<.01	0.019	<.001	<.001	0.457	0.013	<.001
		95% CI	(0.003 - 0.007)	(-0.04 - -0.011)	(0.001 - 0.006)	(0.111 - 0.159)	(0.004 - 0.006)	(-0.019 - 0.009)	(0.0 - 0.002)	(-0.06 - -0.027)
	C2	P-value	0.75	0.753	0.676	0.712	0.644	0.753	0.701	0.702
		P-value	<.01	<.01	0.046	<.001	<.001	0.967	0.28	<.001
		95% CI	(0.002 - 0.006)	(-0.04 - -0.01)	(0.0 - 0.007)	(0.116 - 0.164)	(0.003 - 0.007)	(-0.016 - 0.016)	(-0.001 - 0.002)	(-0.059 - -0.021)
	C3	P-value	0.749	0.752	0.678	0.743	0.647	0.765	0.701	0.712
		P-value	<.001	<.01	0.575	<.001	0.016	0.021	0.022	<.001
		95% CI	(0.004 - 0.005)	(-0.04 - -0.007)	(-0.003 - 0.004)	(0.09 - 0.129)	(0.0 - 0.003)	(-0.022 - -0.002)	(0.0 - 0.003)	(-0.063 - -0.035)
Overall SOFA score increase	R		0.731	0.71	0.673	0.749	0.483	0.574	0.667	0.61
	SpO2	P-value	0.731	0.717	0.675	0.653	0.482	0.569	0.665	0.653
		P-value	0.811	<.01	0.262	<.001	0.29	0.43	0.027	<.001
		95% CI	(-0.001 - 0.001)	(-0.011 - -0.002)	(-0.005 - 0.002)	(0.071 - 0.121)	(-0.001 - 0.002)	(-0.008 - 0.018)	(0.0 - 0.003)	(-0.052 - -0.034)
	C1	P-value	0.731	0.721	0.675	0.622	0.481	0.575	0.665	0.671
		P-value	0.576	0.033	0.387	<.001	0.091	0.904	<.01	<.001
		95% CI	(-0.001 - 0.001)	(-0.02 - -0.001)	(-0.005 - 0.002)	(0.088 - 0.165)	(-0.0 - 0.004)	(-0.021 - 0.019)	(0.001 - 0.004)	(-0.078 - -0.044)
	C2	P-value	0.731	0.722	0.676	0.621	0.483	0.574	0.666	0.668
		P-value	1	0.031	0.209	<.001	0.637	0.978	0.137	<.001
		95% CI	(-0.001 - 0.001)	(-0.022 - -0.001)	(-0.007 - 0.002)	(0.087 - 0.168)	(-0.001 - 0.002)	(-0.016 - 0.016)	(-0.0 - 0.003)	(-0.072 - -0.045)
	C3	P-value	0.73	0.712	0.677	0.635	0.482	0.567	0.664	0.656
		P-value	0.209	0.672	0.064	<.001	0.159	0.317	<.01	<.001
		95% CI	(-0.0 - 0.002)	(-0.009 - 0.006)	(-0.007 - 0.0)	(0.084 - 0.144)	(-0.0 - 0.002)	(-0.008 - 0.023)	(0.002 - 0.006)	(-0.058 - -0.034)

Table B.15: Performance metrics per No/Hidden Hypoxemia groups, in the test set, with RF.

Task	Feature		AUROC		Recall		F1-score		Accuracy	
			No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia
In-hospital mortality	R		0.731	0.746	0.106	0.203	0.181	0.308	0.777	0.701
	SpO2	P-value	0.732	0.746	0.11	0.2	0.187	0.31	0.777	0.707
		P-value	0.492	0.777	0.027	0.76	0.029	0.86	0.147	0.13
		95% CI	(-0.003 - 0.002)	(-0.008 - 0.006)	(-0.007 - -0.001)	(-0.02 - 0.026)	(-0.011 - -0.001)	(-0.03 - 0.025)	(-0.002 - 0.0)	(-0.014 - 0.002)
	C1	P-value	0.73	0.719	0.102	0.176	0.175	0.274	0.779	0.681
		P-value	0.321	0.02	0.032	0.093	0.06	0.088	0.01	0.091
		95% CI	(-0.001 - 0.004)	(0.005 - 0.048)	(0.0 - 0.007)	(-0.005 - 0.059)	(-0.0 - 0.011)	(-0.006 - 0.074)	(-0.004 - -0.001)	(-0.004 - 0.044)
	C2	P-value	0.73	0.716	0.103	0.175	0.176	0.272	0.779	0.683
		P-value	0.177	0.041	0.108	0.039	0.131	0.039	<.01	0.064
		95% CI	(-0.001 - 0.004)	(0.002 - 0.057)	(-0.001 - 0.007)	(0.002 - 0.056)	(-0.002 - 0.011)	(0.002 - 0.069)	(-0.004 - -0.001)	(-0.001 - 0.037)
	C3	P-value	0.733	0.733	0.111	0.206	0.188	0.312	0.777	0.684
		P-value	0.165	0.055	0.025	0.887	0.021	0.806	0.427	0.065
		95% CI	(-0.004 - 0.001)	(-0.0 - 0.026)	(-0.008 - -0.001)	(-0.036 - 0.032)	(-0.013 - -0.001)	(-0.043 - 0.034)	(-0.002 - 0.001)	(-0.001 - 0.035)
Respiratory SOFA score	R		0.771	0.76	0.542	0.721	0.614	0.749	0.724	0.709
	SpO2	P-value	0.765	0.776	0.54	0.64	0.609	0.718	0.72	0.696
		P-value	<.001	<.001	0.275	<.001	0.01	<.001	<.01	0.029
		95% CI	(0.004 - 0.007)	(-0.022 - -0.009)	(-0.002 - 0.006)	(0.067 - 0.097)	(0.001 - 0.009)	(0.021 - 0.042)	(0.002 - 0.007)	(0.002 - 0.024)
	C1	P-value	0.765	0.767	0.532	0.634	0.603	0.721	0.719	0.685
		P-value	<.001	0.294	<.001	<.001	<.001	0.012	<.01	0.02
		95% CI	(0.004 - 0.008)	(-0.021 - 0.007)	(0.007 - 0.012)	(0.068 - 0.107)	(0.008 - 0.014)	(0.008 - 0.048)	(0.003 - 0.008)	(0.005 - 0.043)
	C2	P-value	0.763	0.762	0.53	0.63	0.601	0.715	0.719	0.68
		P-value	<.001	0.804	<.001	<.001	<.001	<.01	<.01	<.01
		95% CI	(0.006 - 0.01)	(-0.014 - 0.011)	(0.008 - 0.015)	(0.073 - 0.11)	(0.009 - 0.016)	(0.016 - 0.052)	(0.003 - 0.008)	(0.011 - 0.046)
	C3	P-value	0.764	0.763	0.537	0.672	0.607	0.738	0.719	0.699
		P-value	<.001	0.548	0.018	<.001	<.01	0.171	<.01	0.211
		95% CI	(0.005 - 0.009)	(-0.013 - 0.008)	(0.001 - 0.008)	(0.032 - 0.067)	(0.003 - 0.01)	(-0.005 - 0.026)	(0.002 - 0.008)	(-0.007 - 0.027)
Overall SOFA score increase	R		0.751	0.725	0.173	0.21	0.276	0.319	0.791	0.684
	SpO2	P-value	0.75	0.736	0.171	0.172	0.274	0.28	0.79	0.69
		P-value	0.259	0.057	0.291	<.001	0.403	<.001	0.541	0.119
		95% CI	(-0.001 - 0.003)	(-0.021 - 0.0)	(-0.002 - 0.006)	(0.028 - 0.048)	(-0.003 - 0.008)	(0.024 - 0.053)	(-0.001 - 0.001)	(-0.012 - 0.002)
	C1	P-value	0.75	0.749	0.173	0.151	0.276	0.251	0.791	0.679
		P-value	0.356	<.01	0.68	<.001	0.765	<.001	0.662	0.24
		95% CI	(-0.001 - 0.003)	(-0.036 - -0.012)	(-0.004 - 0.003)	(0.037 - 0.08)	(-0.005 - 0.004)	(0.039 - 0.096)	(-0.001 - 0.001)	(-0.004 - 0.016)
	C2	P-value	0.75	0.749	0.172	0.153	0.274	0.254	0.791	0.679
		P-value	0.038	<.01	0.671	<.001	0.647	<.001	0.879	0.297
		95% CI	(0.0 - 0.003)	(-0.036 - -0.012)	(-0.004 - 0.006)	(0.038 - 0.076)	(-0.005 - 0.008)	(0.04 - 0.089)	(-0.002 - 0.001)	(-0.006 - 0.017)
	C3	P-value	0.749	0.728	0.171	0.165	0.273	0.268	0.789	0.682
		P-value	0.015	0.578	0.24	<.001	0.187	<.001	0.083	0.593
		95% CI	(0.001 - 0.004)	(-0.014 - 0.008)	(-0.001 - 0.005)	(0.029 - 0.06)	(-0.002 - 0.007)	(0.031 - 0.069)	(-0.0 - 0.003)	(-0.007 - 0.011)

Table B.16: Performance metrics per No/Hidden Hypoxemia groups, in the test set, with XGBoost.

Task	Feature		AUROC		Recall		F1-score		Accuracy	
			No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia	No Hypoxemia	Hidden Hypoxemia
In-hospital mortality	R		0.726	0.735	0.57	0.722	0.476	0.579	0.708	0.652
	SpO2	P-value	0.724	0.728	0.566	0.663	0.472	0.569	0.706	0.667
		P-value	0.377	0.177	0.214	<.01	0.067	0.158	0.166	0.029
		95% CI	(-0.003 - 0.007)	(-0.004 - 0.019)	(-0.003 - 0.012)	(0.027 - 0.091)	(-0.0 - 0.008)	(-0.005 - 0.025)	(-0.001 - 0.006)	(-0.027 - -0.002)
	C1	P-value	0.724	0.699	0.562	0.642	0.471	0.555	0.709	0.649
		P-value	0.494	0.014	0.049	<.001	0.066	0.179	0.742	0.746
		95% CI	(-0.004 - 0.008)	(0.009 - 0.063)	(0.0 - 0.016)	(0.05 - 0.11)	(-0.0 - 0.011)	(-0.013 - 0.061)	(-0.005 - 0.003)	(-0.02 - 0.027)
	C2	P-value	0.725	0.703	0.564	0.655	0.472	0.564	0.71	0.657
		P-value	0.444	0.037	0.049	<.01	0.151	0.489	0.313	0.655
		95% CI	(-0.003 - 0.006)	(0.002 - 0.061)	(0.0 - 0.012)	(0.023 - 0.112)	(-0.002 - 0.009)	(-0.031 - 0.061)	(-0.005 - 0.002)	(-0.03 - 0.02)
	C3	P-value	0.726	0.722	0.569	0.689	0.477	0.59	0.709	0.664
		P-value	0.957	0.047	0.645	0.025	0.855	0.237	0.569	0.021
		95% CI	(-0.004 - 0.004)	(0.0 - 0.027)	(-0.005 - 0.007)	(0.005 - 0.06)	(-0.005 - 0.004)	(-0.03 - 0.008)	(-0.003 - 0.002)	(-0.021 - -0.002)
Respiratory SOFA score	R		0.77	0.761	0.65	0.843	0.65	0.764	0.717	0.686
	SpO2	P-value	0.766	0.775	0.65	0.754	0.647	0.759	0.713	0.71
		P-value	<.01	<.01	0.78	<.001	0.051	0.269	<.01	0.01
		95% CI	(0.002 - 0.007)	(-0.022 - -0.006)	(-0.005 - 0.004)	(0.077 - 0.102)	(-0.0 - 0.007)	(-0.005 - 0.017)	(0.001 - 0.007)	(-0.041 - -0.007)
	C1	P-value	0.764	0.767	0.647	0.737	0.644	0.761	0.713	0.701
		P-value	<.001	0.44	0.211	<.001	<.001	0.681	<.001	0.192
		95% CI	(0.004 - 0.007)	(-0.023 - 0.011)	(-0.002 - 0.006)	(0.093 - 0.119)	(0.004 - 0.009)	(-0.016 - 0.023)	(0.002 - 0.006)	(-0.04 - 0.009)
	C2	P-value	0.764	0.766	0.642	0.728	0.64	0.76	0.712	0.706
		P-value	<.001	0.523	<.01	<.001	<.001	0.6	<.001	0.096
		95% CI	(0.005 - 0.008)	(-0.022 - 0.012)	(0.003 - 0.012)	(0.096 - 0.133)	(0.007 - 0.012)	(-0.015 - 0.025)	(0.004 - 0.007)	(-0.043 - 0.004)
	C3	P-value	0.764	0.76	0.65	0.771	0.645	0.765	0.711	0.7
		P-value	<.001	0.937	0.794	<.001	<.001	0.93	<.001	0.144
		95% CI	(0.005 - 0.008)	(-0.011 - 0.012)	(-0.004 - 0.003)	(0.056 - 0.088)	(0.003 - 0.007)	(-0.016 - 0.014)	(0.004 - 0.008)	(-0.035 - 0.006)
Overall SOFA score increase	R		0.76	0.739	0.631	0.807	0.511	0.601	0.721	0.623
	SpO2	P-value	0.759	0.754	0.632	0.7	0.507	0.601	0.716	0.674
		P-value	0.126	0.024	0.623	<.001	<.01	1	<.001	<.001
		95% CI	(-0.0 - 0.003)	(-0.028 - -0.003)	(-0.005 - 0.003)	(0.082 - 0.133)	(0.002 - 0.006)	(-0.014 - 0.014)	(0.003 - 0.008)	(-0.063 - -0.038)
	C1	P-value	0.759	0.758	0.632	0.66	0.509	0.607	0.719	0.694
		P-value	0.225	0.011	0.427	<.001	0.246	0.59	0.049	<.001
		95% CI	(-0.001 - 0.002)	(-0.033 - -0.006)	(-0.004 - 0.002)	(0.096 - 0.197)	(-0.002 - 0.005)	(-0.031 - 0.019)	(0.0 - 0.005)	(-0.087 - -0.056)
	C2	P-value	0.758	0.759	0.628	0.652	0.507	0.603	0.718	0.691
		P-value	0.014	0.015	0.16	<.001	0.016	0.88	<.01	<.001
		95% CI	(0.001 - 0.004)	(-0.036 - -0.005)	(-0.001 - 0.007)	(0.121 - 0.189)	(0.001 - 0.007)	(-0.034 - 0.03)	(0.001 - 0.005)	(-0.095 - -0.04)
	C3	P-value	0.757	0.737	0.633	0.677	0.508	0.582	0.716	0.656
		P-value	<.01	0.665	0.458	<.001	0.154	0.034	<.01	<.001
		95% CI	(0.001 - 0.005)	(-0.008 - 0.012)	(-0.008 - 0.004)	(0.095 - 0.166)	(-0.001 - 0.008)	(0.002 - 0.036)	(0.002 - 0.008)	(-0.047 - -0.018)

Table B.17: Fairness metrics in the test set, in the Blood-gas and Oximetry study.

Task	Feature		**LR**		**RF**		**XGBoost**	
			Demographic parity difference	Equalized odds ratio	Demographic parity difference	Equalized odds ratio	Demographic parity difference	Equalized odds ratio
In-hospital mortality	Reference		0.103	0.642	0.047	0.212	0.127	0.653
	SpO2		0.108	0.629	0.044	0.19	0.129	0.633
		P-value	0.438	0.542	0.455	0.289	0.789	0.32
		95% CI	(-0.019 - 0.009)	(-0.034 - 0.06)	(-0.007 - 0.014)	(-0.022 - 0.067)	(-0.021 - 0.017)	(-0.022 - 0.061)
	Correction 1		0.103	0.658	0.042	0.171	0.12	0.67
		P-value	0.81	0.43	0.367	0.436	0.61	0.425
		95% CI	(-0.008 - 0.007)	(-0.06 - 0.028)	(-0.008 - 0.019)	(-0.073 - 0.156)	(-0.021 - 0.034)	(-0.065 - 0.03)
	Correction 2		0.111	0.602	0.044	0.181	0.119	0.656
		P-value	0.312	0.109	0.506	0.41	0.475	0.873
		95% CI	(-0.027 - 0.01)	(-0.011 - 0.089)	(-0.008 - 0.015)	(-0.05 - 0.112)	(-0.015 - 0.031)	(-0.05 - 0.043)
	Correction 3		0.105	0.66	0.04	0.182	0.122	0.663
		P-value	0.555	0.264	0.204	0.533	0.607	0.647
		95% CI	(-0.009 - 0.005)	(-0.052 - 0.016)	(-0.005 - 0.019)	(-0.074 - 0.134)	(-0.016 - 0.026)	(-0.06 - 0.039)
Respiratory SOFA score	Reference		0.117	0.752	0.127	0.604	0.116	0.704
	SpO2		0.112	0.74	0.129	0.596	0.128	0.689
		P-value	0.221	0.324	0.821	0.656	0.05	0.413
		95% CI	(-0.003 - 0.013)	(-0.014 - 0.039)	(-0.015 - 0.012)	(-0.031 - 0.047)	(-0.024 - 0.0)	(-0.024 - 0.053)
	Correction 1		0.116	0.742	0.127	0.596	0.129	0.707
		P-value	0.912	0.542	0.965	0.733	0.091	0.891
		95% CI	(-0.013 - 0.015)	(-0.028 - 0.049)	(-0.015 - 0.015)	(-0.041 - 0.056)	(-0.03 - 0.003)	(-0.053 - 0.047)
	Correction 2		0.119	0.73	0.123	0.631	0.13	0.691
		P-value	0.712	0.229	0.57	0.236	0.138	0.613
		95% CI	(-0.014 - 0.01)	(-0.017 - 0.061)	(-0.012 - 0.02)	(-0.075 - 0.021)	(-0.033 - 0.005)	(-0.042 - 0.067)
	Correction 3		0.117	0.74	0.132	0.584	0.13	0.702
		P-value	0.987	0.511	0.616	0.439	0.068	0.937
		95% CI	(-0.013 - 0.013)	(-0.028 - 0.052)	(-0.028 - 0.018)	(-0.035 - 0.074)	(-0.03 - 0.001)	(-0.04 - 0.043)
Overall SOFA score increase	Reference		0.089	0.758	0.03	0.456	0.078	0.769
	SpO2		0.094	0.75	0.029	0.409	0.073	0.753
		P-value	0.464	0.56	0.793	0.223	0.31	0.137
		95% CI	(-0.019 - 0.009)	(-0.025 - 0.043)	(-0.004 - 0.006)	(-0.034 - 0.127)	(-0.005 - 0.013)	(-0.006 - 0.04)
	Correction 1		0.095	0.756	0.028	0.468	0.082	0.736
		P-value	0.207	0.716	0.353	0.769	0.602	0.136
		95% CI	(-0.015 - 0.004)	(-0.014 - 0.02)	(-0.002 - 0.006)	(-0.099 - 0.076)	(-0.022 - 0.014)	(-0.012 - 0.078)
	Correction 2		0.093	0.76	0.033	0.391	0.082	0.729
		P-value	0.463	0.879	0.285	0.019	0.677	0.095
		95% CI	(-0.014 - 0.007)	(-0.025 - 0.021)	(-0.009 - 0.003)	(0.013 - 0.117)	(-0.028 - 0.019)	(-0.009 - 0.09)
	Correction 3		0.096	0.746	0.032	0.411	0.082	0.74
		P-value	0.222	0.269	0.331	0.327	0.489	0.095
		95% CI	(-0.018 - 0.005)	(-0.012 - 0.037)	(-0.008 - 0.003)	(-0.054 - 0.144)	(-0.02 - 0.01)	(-0.006 - 0.065)

B.2 Thermometry

B.2.1 Training set

Table B.18: Performance metrics in the training set per race and ethnicity groups, in the Thermometry study.

Logistic Regression

Task	Feature	AUROC						Recall						F1-score						Accuracy					
		All	A	B	HL	O	W	All	A	B	HL	O	W	All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	Reference	0.826	0.920	0.853	0.864	0.823	0.814	0.748	0.881	0.778	0.834	0.713	0.740	0.577	0.672	0.549	0.703	0.651	0.544	0.741	0.778	0.743	0.723	0.733	0.741
	Temporal	0.828	0.907	0.851	0.855	0.830	0.815	0.751	0.881	0.761	0.818	0.719	0.745	0.581	0.718	0.542	0.699	0.657	0.547	0.744	0.822	0.742	0.722	0.737	0.742
	Correction 1	0.825	0.916	0.853	0.861	0.825	0.812	0.749	0.881	0.770	0.813	0.716	0.742	0.577	0.689	0.545	0.694	0.654	0.542	0.740	0.796	0.742	0.718	0.735	0.739
	Correction 2	0.825	0.917	0.854	0.862	0.826	0.812	0.747	0.881	0.765	0.810	0.717	0.740	0.576	0.681	0.547	0.686	0.655	0.543	0.740	0.787	0.746	0.709	0.736	0.740
	Correction 3	0.826	0.918	0.852	0.864	0.823	0.814	0.747	0.881	0.772	0.825	0.712	0.739	0.576	0.685	0.543	0.697	0.652	0.542	0.740	0.791	0.739	0.717	0.734	0.740
Respiratory SOFA score	Reference	0.883	0.660	0.927	0.875	0.854	0.889	0.773	0.637	0.862	0.798	0.762	0.764	0.756	0.489	0.846	0.807	0.741	0.755	0.837	0.657	0.889	0.788	0.776	0.855
	Temporal	0.882	0.645	0.925	0.874	0.858	0.889	0.766	0.546	0.857	0.788	0.763	0.758	0.752	0.439	0.840	0.804	0.739	0.752	0.835	0.640	0.885	0.788	0.774	0.853
	Correction 1	0.885	0.665	0.928	0.876	0.857	0.891	0.773	0.631	0.882	0.754	0.765	0.765	0.755	0.472	0.857	0.781	0.743	0.755	0.836	0.635	0.896	0.766	0.777	0.854
	Correction 2	0.885	0.664	0.930	0.875	0.857	0.891	0.773	0.618	0.879	0.763	0.761	0.766	0.755	0.468	0.854	0.788	0.742	0.755	0.836	0.637	0.894	0.772	0.777	0.854
	Correction 3	0.884	0.666	0.927	0.879	0.851	0.890	0.772	0.618	0.868	0.774	0.757	0.766	0.756	0.470	0.850	0.794	0.737	0.758	0.837	0.641	0.891	0.776	0.773	0.857
Overall SOFA score increase	Reference	0.733	0.729	0.750	0.722	0.749	0.730	0.654	0.600	0.766	0.706	0.605	0.650	0.513	0.538	0.472	0.630	0.519	0.507	0.674	0.603	0.609	0.652	0.677	0.688
	Temporal	0.732	0.741	0.755	0.723	0.749	0.729	0.653	0.627	0.776	0.708	0.608	0.646	0.510	0.551	0.473	0.627	0.522	0.502	0.670	0.607	0.604	0.645	0.680	0.683
	Correction 1	0.732	0.732	0.752	0.715	0.746	0.730	0.655	0.649	0.759	0.711	0.600	0.650	0.514	0.569	0.475	0.630	0.514	0.506	0.674	0.622	0.617	0.648	0.673	0.686
	Correction 2	0.732	0.727	0.749	0.718	0.750	0.731	0.651	0.587	0.753	0.682	0.606	0.652	0.509	0.531	0.460	0.609	0.518	0.506	0.671	0.600	0.598	0.630	0.674	0.686
	Correction 3	0.732	0.728	0.749	0.722	0.750	0.731	0.651	0.584	0.768	0.698	0.609	0.647	0.511	0.525	0.466	0.630	0.522	0.506	0.673	0.590	0.598	0.655	0.678	0.689

Random Forest

In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

XGBoost

In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table B.19: Performance metrics in the training set per disparity groups, in the Thermometry study.

Logistic Regression

Task	Feature	AUROC				Recall				F1-score				Accuracy			
		(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)
In-hospital mortality	Reference	0.797	0.840	0.806	0.876	0.607	0.770	0.725	0.826	0.582	0.612	0.527	0.644	0.788	0.764	0.716	0.722
	Temporal	0.794	0.840	0.806	0.876	0.674	0.787	0.712	0.794	0.594	0.613	0.527	0.672	0.776	0.760	0.721	0.765
	Correction 1	0.831	0.832	0.828	0.729	0.772	0.793	0.747	0.604	0.619	0.577	0.576	0.554	0.725	0.716	0.752	0.607
	Correction 2	0.801	0.833	0.827	0.713	0.739	0.756	0.756	0.596	0.577	0.572	0.584	0.540	0.685	0.748	0.747	0.572
	Correction 3		0.825	0.826		0.000	0.732	0.761	0.000	0.000	0.572	0.580	0.000		0.744	0.737	
Respiratory SOFA score	Reference	0.890	0.892	0.875	0.847	0.685	0.801	0.755	0.750	0.698	0.794	0.721	0.739	0.828	0.849	0.837	0.790
	Temporal	0.890	0.894	0.874	0.843	0.730	0.809	0.738	0.689	0.687	0.794	0.717	0.716	0.806	0.847	0.837	0.784
	Correction 1	0.896	0.908	0.879	0.806	0.885	0.813	0.758	0.706	0.811	0.817	0.737	0.676	0.832	0.864	0.829	0.836
	Correction 2	0.875	0.882	0.890	0.801	0.837	0.769	0.774	0.692	0.804	0.765	0.743	0.681	0.823	0.844	0.828	0.839
	Correction 3		0.882	0.886		0.000	0.764	0.780	0.000	0.000	0.754	0.758	0.000		0.834	0.840	
Overall SOFA score increase	Reference	0.744	0.720	0.748	0.718	0.767	0.621	0.680	0.615	0.599	0.485	0.510	0.581	0.700	0.654	0.692	0.664
	Temporal	0.747	0.719	0.745	0.719	0.769	0.612	0.682	0.639	0.616	0.480	0.505	0.582	0.720	0.652	0.685	0.652
	Correction 1	0.713	0.760	0.724	0.753	0.473	0.675	0.652	0.737	0.421	0.537	0.508	0.568	0.773	0.700	0.663	0.679
	Correction 2	0.716	0.728	0.736	0.755	0.519	0.640	0.664	0.751	0.475	0.498	0.517	0.572	0.759	0.666	0.667	0.675
	Correction 3		0.732	0.732		0.000	0.662	0.640	0.000	0.000	0.516	0.506	0.000		0.672	0.675	

Random Forest

Task	Feature	AUROC				Recall				F1-score				Accuracy			
		(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)
In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3		1.000	1.000		0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000		1.000	1.000	
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3		1.000	1.000		0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000		1.000	1.000	
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3		1.000	1.000		0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000		1.000	1.000	

XGBoost

Task	Feature	AUROC				Recall				F1-score				Accuracy			
		(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)
In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3		1.000	1.000		0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000		1.000	1.000	
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3		1.000	1.000		0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000		1.000	1.000	
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	0.999
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	1.000	1.000	1.000	0.997
	Correction 3		1.000	1.000		0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000		1.000	1.000	

Table B.20: Performance metrics in the training set by hidden hypothermia, normothermia and hidden fever, in the Thermometry study.

Logistic Regression

Task	Feature	AUROC			Recall			F1-score			Accuracy		
		Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever
In-hospital mortality	Reference	0.936	0.823	0.802	0.904	0.756	0.560	0.820	0.549	0.552	0.815	0.744	0.761
	Temporal	0.937	0.825	0.800	0.902	0.759	0.605	0.838	0.552	0.556	0.837	0.746	0.744
	Correction 1	1.000	0.823	0.783	0.980	0.767	0.641	0.989	0.541	0.486	0.985	0.740	0.719
	Correction 2	1.000	0.824	0.784	0.989	0.765	0.641	0.994	0.544	0.487	0.992	0.742	0.719
	Correction 3	1.000	0.822	0.968	0.700	0.760	0.567	0.700	0.551	0.707	1.000	0.737	0.813
Respiratory SOFA score	Reference	0.927	0.882	0.883	0.950	0.770	0.747	0.919	0.739	0.727	0.910	0.852	0.819
	Temporal	0.923	0.883	0.881	0.922	0.762	0.798	0.906	0.734	0.723	0.898	0.849	0.803
	Correction 1	0.991	0.886	0.844	1.000	0.760	0.804	0.921	0.743	0.774	0.928	0.843	0.788
	Correction 2	0.987	0.887	0.853	1.000	0.763	0.803	0.921	0.745	0.780	0.927	0.844	0.794
	Correction 3		0.887	0.973	0.000	0.770	0.608	0.000	0.750	0.689	0.902	0.846	0.759
Overall SOFA score increase	Reference	0.826	0.714	0.688	0.782	0.624	0.517	0.725	0.467	0.364	0.744	0.655	0.733
	Temporal	0.828	0.713	0.683	0.790	0.621	0.525	0.726	0.464	0.383	0.744	0.652	0.750
	Correction 1	0.691	0.730	0.591	0.500	0.655	0.400	0.464	0.485	0.256	0.766	0.660	0.781
	Correction 2	0.614	0.731	0.587	0.367	0.652	0.400	0.320	0.483	0.256	0.757	0.657	0.780
	Correction 3	1.000	0.728	0.808	1.000	0.649	1.000	1.000	0.488	0.791	1.000	0.658	0.786

Random Forest

In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	0.700	1.000	1.000	0.700	1.000	1.000	1.000	1.000	1.000
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

XGBoost

In-hospital mortality	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	0.700	1.000	1.000	0.700	1.000	1.000	1.000	1.000	1.000
Respiratory SOFA score	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3		1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
Overall SOFA score increase	Reference	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Temporal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correction 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table B.21: Fairness metrics in the training set, in the Thermometry study.

Task	Feature	LR		RF		XGBoost	
		Demographic parity difference	Equalized odds ratio	Demographic parity difference	Equalized odds ratio	Demographic parity difference	Equalized odds ratio
In-hospital mortality	Reference	0.188	0.646	0.196	1.000	0.196	1.000
	Temporal	0.184	0.562	0.196	1.000	0.196	1.000
	Correction 1	0.178	0.627	0.196	1.000	0.196	1.000
	Correction 2	0.188	0.620	0.196	1.000	0.196	1.000
	Correction 3	0.187	0.596	0.196	1.000	0.196	1.000
Respiratory SOFA score	Reference	0.248	0.291	0.305	1.000	0.305	1.000
	Temporal	0.244	0.291	0.305	1.000	0.305	1.000
	Correction 1	0.226	0.261	0.305	1.000	0.305	1.000
	Correction 2	0.229	0.275	0.305	1.000	0.305	1.000
	Correction 3	0.233	0.272	0.305	1.000	0.305	1.000
Overall SOFA score increase	Reference	0.162	0.628	0.202	1.000	0.202	0.900
	Temporal	0.181	0.600	0.202	1.000	0.202	0.900
	Correction 1	0.166	0.651	0.202	1.000	0.202	0.900
	Correction 2	0.174	0.624	0.202	1.000	0.202	0.800
	Correction 3	0.176	0.613	0.202	1.000	0.202	1.000

B.2.2 Test set

Table B.22: AUROC and recall in the test set with LR, in the Thermometry study.

Task	Feature		AUROC						Recall					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.798	0.904	0.784	0.805	0.769	0.789	0.705	0.652	0.678	0.541	0.694	0.685
	T		0.801	0.904	0.760	0.804	0.767	0.793	0.717	0.652	0.678	0.527	0.693	0.712
		P-value	0.145		0.364	0.363	0.820	0.131	0.218			0.343	0.869	0.111
		95% CI	(-0.008 - 0.001)	(0.0 - 0.0)	(-0.033 - 0.08)	(-0.002 - 0.005)	(-0.013 - 0.016)	(-0.009 - 0.001)	(-0.035 - 0.009)	(0.0 - 0.0)	(0.0 - 0.0)	(-0.018 - 0.047)	(-0.015 - 0.017)	(-0.062 - 0.008)
	C1		0.796	0.894	0.744	0.817	0.770	0.786	0.710	0.652	0.566	0.527	0.696	0.700
		P-value	0.085	0.363	0.125	0.249	0.595	0.127	0.604		0.287	0.343	0.828	0.284
		95% CI	(-0.0 - 0.004)	(-0.015 - 0.035)	(-0.014 - 0.095)	(-0.036 - 0.012)	(-0.007 - 0.004)	(-0.001 - 0.007)	(-0.026 - 0.016)	(0.0 - 0.0)	(-0.112 - 0.337)	(-0.018 - 0.047)	(-0.026 - 0.021)	(-0.047 - 0.015)
	C2		0.797	0.904	0.743	0.816	0.765	0.786	0.705	0.652	0.566	0.527	0.676	0.700
		P-value	0.433		0.134	0.298	0.255	0.110	1.000		0.287	0.343	0.179	0.275
		95% CI	(-0.001 - 0.003)	(0.0 - 0.0)	(-0.016 - 0.097)	(-0.035 - 0.013)	(-0.003 - 0.011)	(-0.001 - 0.006)	(-0.018 - 0.018)	(0.0 - 0.0)	(-0.112 - 0.337)	(-0.018 - 0.047)	(-0.01 - 0.046)	(-0.046 - 0.015)
	C3		0.798	0.894	0.777	0.810	0.764	0.790	0.714	0.652	0.678	0.527	0.694	0.704
		P-value	0.708	0.363	0.297	0.363	0.374	0.323	0.133			0.343		0.087
		95% CI	(-0.005 - 0.003)	(-0.015 - 0.035)	(-0.007 - 0.02)	(-0.017 - 0.007)	(-0.006 - 0.014)	(-0.005 - 0.002)	(-0.023 - 0.004)	(0.0 - 0.0)	(0.0 - 0.0)	(-0.018 - 0.047)	(0.0 - 0.0)	(-0.043 - 0.004)
Respiratory SOFA score	R		0.866	0.555	0.930	0.774	0.828	0.877	0.758	0.400	0.849	0.436	0.781	0.755
	T		0.861	0.549	0.936	0.791	0.833	0.876	0.742	0.327	0.849	0.461	0.755	0.740
		P-value	0.123	0.932	0.602	0.507	0.357	0.835	0.092	0.170		0.343	0.287	0.443
		95% CI	(-0.002 - 0.011)	(-0.161 - 0.173)	(-0.03 - 0.018)	(-0.071 - 0.039)	(-0.017 - 0.007)	(-0.013 - 0.016)	(-0.003 - 0.035)	(-0.038 - 0.184)	(0.0 - 0.0)	(-0.082 - 0.032)	(-0.026 - 0.079)	(-0.027 - 0.056)
	C1		0.865	0.592	0.941	0.772	0.828	0.874	0.745	0.360	0.820	0.496	0.796	0.731
		P-value	0.858	0.618	0.163	0.351	0.990	0.653	0.246	0.343	0.240	0.260	0.223	0.351
		95% CI	(-0.01 - 0.012)	(-0.213 - 0.14)	(-0.028 - 0.005)	(-0.003 - 0.008)	(-0.018 - 0.018)	(-0.014 - 0.021)	(-0.011 - 0.039)	(-0.05 - 0.13)	(-0.023 - 0.08)	(-0.173 - 0.053)	(-0.042 - 0.011)	(-0.031 - 0.079)
	C2		0.862	0.570	0.934	0.786	0.833	0.874	0.740	0.400	0.814	0.486	0.754	0.735
		P-value	0.299	0.855	0.135	0.351	0.502	0.500	0.175		0.320	0.343	0.344	0.343
		95% CI	(-0.004 - 0.012)	(-0.218 - 0.187)	(-0.011 - 0.002)	(-0.038 - 0.016)	(-0.019 - 0.01)	(-0.008 - 0.016)	(-0.01 - 0.047)	(0.0 - 0.0)	(-0.04 - 0.109)	(-0.163 - 0.063)	(-0.034 - 0.087)	(-0.025 - 0.066)
	C3		0.858	0.567	0.937	0.800	0.836	0.868	0.738	0.380	0.814	0.486	0.774	0.729
		P-value	0.103	0.861	0.211	0.173	0.398	0.166	0.122	0.343	0.320	0.343	0.769	0.329
		95% CI	(-0.002 - 0.018)	(-0.177 - 0.153)	(-0.02 - 0.005)	(-0.065 - 0.014)	(-0.025 - 0.011)	(-0.005 - 0.025)	(-0.007 - 0.048)	(-0.025 - 0.065)	(-0.04 - 0.109)	(-0.163 - 0.063)	(-0.047 - 0.062)	(-0.031 - 0.083)
Overall SOFA score increase	R		0.688	0.696	0.662	0.749	0.719	0.688	0.593	0.513	0.617	0.396	0.577	0.586
	T		0.690	0.696	0.671	0.753	0.711	0.690	0.601	0.513	0.612	0.427	0.563	0.595
		P-value	0.357		0.518	0.896	0.385	0.557	0.196		0.893	0.343	0.686	0.448
		95% CI	(-0.006 - 0.002)	(0.0 - 0.0)	(-0.038 - 0.021)	(-0.071 - 0.064)	(-0.011 - 0.027)	(-0.005 - 0.003)	(-0.022 - 0.005)	(0.0 - 0.0)	(-0.065 - 0.073)	(-0.1 - 0.039)	(-0.061 - 0.088)	(-0.035 - 0.017)
	C1		0.687	0.704	0.672	0.748	0.710	0.690	0.594	0.513	0.615	0.435	0.563	0.583
		P-value	0.699	0.356	0.520	0.979	0.096	0.380	0.574		0.934	0.343	0.343	0.833
		95% CI	(-0.005 - 0.007)	(-0.028 - 0.012)	(-0.04 - 0.022)	(-0.075 - 0.077)	(-0.002 - 0.02)	(-0.007 - 0.003)	(-0.007 - 0.004)	(0.0 - 0.0)	(-0.043 - 0.047)	(-0.125 - 0.048)	(-0.017 - 0.045)	(-0.023 - 0.027)
	C2		0.689	0.695	0.663	0.754	0.722	0.695	0.594	0.513	0.600	0.404	0.572	0.590
		P-value	0.660	0.356	0.913	0.863	0.625	<.01	0.732		0.551	0.343	0.343	0.629
		95% CI	(-0.007 - 0.005)	(-0.002 - 0.005)	(-0.015 - 0.013)	(-0.08 - 0.069)	(-0.018 - 0.011)	(-0.011 - -0.002)	(-0.01 - 0.008)	(0.0 - 0.0)	(-0.044 - 0.078)	(-0.025 - 0.01)	(-0.006 - 0.015)	(-0.024 - 0.016)
	C3		0.692	0.712	0.665	0.709	0.724	0.694	0.584	0.513	0.617	0.454	0.572	0.569
		P-value	0.193	0.187	0.775	0.205	0.620	0.082	0.218		1.000	0.277	0.343	0.260
		95% CI	(-0.011 - 0.002)	(-0.041 - 0.01)	(-0.02 - 0.015)	(-0.028 - 0.107)	(-0.026 - 0.017)	(-0.012 - 0.001)	(-0.006 - 0.023)	(0.0 - 0.0)	(-0.048 - 0.048)	(-0.17 - 0.055)	(-0.006 - 0.015)	(-0.015 - 0.048)

Table B.23: F1-score and accuracy in the test set with LR, in the Thermometry study.

Task	Feature		F1-score						Accuracy					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.546	0.517	0.490	0.490	0.593	0.502	0.724	0.815	0.728	0.727	0.705	0.719
	T	P-value	0.556	0.540	0.463	0.481	0.597	0.518	0.730	0.829	0.710	0.712	0.708	0.727
		95% CI	0.125 (-0.023 - 0.003)	0.343 (-0.075 - 0.029)	0.343 (-0.034 - 0.087)	0.343 (-0.011 - 0.029)	0.504 (-0.018 - 0.01)	0.056 (-0.032 - 0.001)	0.155 (-0.015 - 0.003)	0.347 (-0.049 - 0.019)	0.400 (-0.029 - 0.065)	0.343 (-0.018 - 0.047)	0.494 (-0.015 - 0.008)	0.044 (-0.016 - -0.0)
	C1	P-value	0.550	0.524	0.412	0.481	0.599	0.512	0.726	0.824	0.684	0.712	0.709	0.724
		95% CI	0.477 (-0.017 - 0.009)	0.343 (-0.023 - 0.009)	0.272 (-0.072 - 0.226)	0.343 (-0.011 - 0.029)	0.487 (-0.028 - 0.014)	0.205 (-0.027 - 0.007)	0.497 (-0.008 - 0.004)	0.347 (-0.03 - 0.012)	0.137 (-0.017 - 0.104)	0.343 (-0.018 - 0.047)	0.625 (-0.026 - 0.016)	0.202 (-0.013 - 0.003)
	C2	P-value	0.548	0.535	0.417	0.481	0.585	0.512	0.725	0.825	0.690	0.712	0.698	0.724
		95% CI	0.771 (-0.015 - 0.012)	0.489 (-0.072 - 0.037)	0.302 (-0.078 - 0.224)	0.343 (-0.011 - 0.029)	0.409 (-0.012 - 0.026)	0.174 (-0.026 - 0.005)	0.776 (-0.01 - 0.007)	0.557 (-0.047 - 0.027)	0.212 (-0.026 - 0.101)	0.343 (-0.018 - 0.047)	0.402 (-0.01 - 0.023)	0.254 (-0.015 - 0.004)
	C3	P-value	0.550	0.507	0.486	0.481	0.594	0.511	0.725	0.805	0.726	0.712	0.708	0.722
		95% CI	0.245 (-0.012 - 0.004)	0.343 (-0.013 - 0.033)	0.343 (-0.005 - 0.012)	0.343 (-0.011 - 0.029)	0.780 (-0.009 - 0.007)	0.098 (-0.021 - 0.002)	0.639 (-0.006 - 0.004)	0.347 (-0.013 - 0.032)	0.775 (-0.013 - 0.017)	0.343 (-0.018 - 0.047)	0.490 (-0.013 - 0.007)	0.275 (-0.01 - 0.003)
Respiratory SOFA score	R		0.736	0.351	0.786	0.465	0.717	0.734	0.822	0.746	0.860	0.648	0.754	0.839
	T	P-value	0.726	0.299	0.764	0.480	0.718	0.726	0.815	0.679	0.846	0.660	0.756	0.836
		95% CI	0.173 (-0.006 - 0.027)	0.132 (-0.019 - 0.123)	0.389 (-0.033 - 0.076)	0.343 (-0.047 - 0.018)	0.962 (-0.056 - 0.054)	0.440 (-0.015 - 0.032)	0.213 (-0.005 - 0.018)	0.104 (-0.017 - 0.151)	0.408 (-0.022 - 0.049)	0.343 (-0.036 - 0.014)	0.901 (-0.043 - 0.038)	0.509 (-0.008 - 0.015)
	C1	P-value	0.728	0.333	0.782	0.498	0.736	0.718	0.819	0.740	0.860	0.655	0.771	0.833
		95% CI	0.293 (-0.008 - 0.023)	0.539 (-0.047 - 0.084)	0.824 (-0.033 - 0.04)	0.303 (-0.1 - 0.035)	0.194 (-0.049 - 0.011)	0.292 (-0.017 - 0.049)	0.362 (-0.005 - 0.012)	0.737 (-0.033 - 0.045)	0.989 (-0.031 - 0.031)	0.628 (-0.035 - 0.022)	0.224 (-0.046 - 0.012)	0.344 (-0.008 - 0.02)
	C2	P-value	0.728	0.362	0.774	0.495	0.705	0.724	0.820	0.749	0.849	0.660	0.749	0.838
		95% CI	0.364 (-0.011 - 0.026)	0.222 (-0.03 - 0.008)	0.543 (-0.03 - 0.054)	0.343 (-0.098 - 0.038)	0.625 (-0.041 - 0.064)	0.490 (-0.021 - 0.041)	0.608 (-0.008 - 0.013)	0.909 (-0.054 - 0.049)	0.507 (-0.026 - 0.048)	0.343 (-0.036 - 0.014)	0.773 (-0.032 - 0.041)	0.811 (-0.012 - 0.015)
	C3	P-value	0.724	0.342	0.767	0.495	0.713	0.721	0.817	0.730	0.851	0.660	0.751	0.835
		95% CI	0.169 (-0.006 - 0.03)	0.343 (-0.012 - 0.031)	0.500 (-0.041 - 0.078)	0.343 (-0.098 - 0.038)	0.805 (-0.036 - 0.045)	0.434 (-0.023 - 0.049)	0.243 (-0.004 - 0.015)	0.343 (-0.021 - 0.054)	0.667 (-0.038 - 0.057)	0.343 (-0.036 - 0.014)	0.793 (-0.022 - 0.028)	0.602 (-0.012 - 0.02)
Overall SOFA score increase	R		0.465	0.392	0.363	0.302	0.481	0.460	0.643	0.589	0.500	0.592	0.664	0.666
	T	P-value	0.466	0.392	0.354	0.327	0.468	0.465	0.641	0.569	0.485	0.616	0.648	0.666
		95% CI	0.779 (-0.013 - 0.01)	(0.0 - 0.0)	0.631 (-0.032 - 0.05)	0.290 (-0.075 - 0.025)	0.546 (-0.034 - 0.06)	0.602 (-0.024 - 0.015)	0.649 (-0.008 - 0.012)	0.343 (-0.025 - 0.065)	0.321 (-0.017 - 0.047)	0.241 (-0.067 - 0.019)	0.119 (-0.005 - 0.036)	0.965 (-0.015 - 0.015)
	C1	P-value	0.466	0.392	0.361	0.322	0.471	0.459	0.643	0.589	0.500	0.607	0.658	0.666
		95% CI	0.695 (-0.008 - 0.006)	(0.0 - 0.0)	0.891 (-0.02 - 0.022)	0.495 (-0.085 - 0.044)	0.305 (-0.011 - 0.03)	0.904 (-0.014 - 0.015)	0.920 (-0.009 - 0.008)	(0.0 - 0.0)	(0.0 - 0.0)	0.591 (-0.074 - 0.045)	0.522 (-0.013 - 0.025)	0.941 (-0.012 - 0.012)
	C2	P-value	0.462	0.392	0.356	0.308	0.463	0.461	0.638	0.569	0.497	0.597	0.646	0.664
		95% CI	0.343 (-0.004 - 0.009)	(0.0 - 0.0)	0.695 (-0.033 - 0.047)	0.343 (-0.021 - 0.008)	0.045 (0.0 - 0.035)	0.846 (-0.014 - 0.011)	0.094 (-0.001 - 0.011)	0.343 (-0.025 - 0.065)	0.857 (-0.037 - 0.043)	0.343 (-0.015 - 0.006)	0.044 (0.001 - 0.035)	0.707 (-0.009 - 0.013)
	C3	P-value	0.465	0.392	0.377	0.348	0.478	0.455	0.647	0.656	0.528	0.617	0.668	0.665
		95% CI	0.960 (-0.013 - 0.013)	(0.0 - 0.0)	0.261 (-0.041 - 0.012)	0.274 (-0.136 - 0.044)	0.699 (-0.016 - 0.023)	0.582 (-0.016 - 0.027)	0.492 (-0.019 - 0.01)	0.343 (-0.218 - 0.084)	0.025 (-0.052 - -0.004)	0.249 (-0.07 - 0.021)	0.721 (-0.034 - 0.024)	0.940 (-0.017 - 0.018)

Table B.24: AUROC and recall in the test set with RF, in the Thermometry study.

Task	Feature		AUROC						Recall					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.803	0.911	0.835	0.825	0.788	0.786	0.243	0.229	0.231	0.260	0.307	0.191
	T		0.784	0.816	0.775	0.779	0.798	0.775	0.268	0.314	0.243	0.246	0.357	0.204
		P-value	0.012	0.253	0.075	0.363	0.666	0.188	0.162	0.255	0.343	0.343	0.360	0.666
		95% CI	(0.005 - 0.032)	(-0.095 - 0.287)	(-0.008 - 0.129)	(-0.071 - 0.162)	(-0.064 - 0.043)	(-0.006 - 0.027)	(-0.064 - 0.012)	(-0.245 - 0.074)	(-0.041 - 0.016)	(-0.018 - 0.047)	(-0.168 - 0.067)	(-0.083 - 0.055)
	C1		0.798	0.874	0.830	0.756	0.794	0.781	0.275	0.293	0.251	0.246	0.340	0.223
		P-value	0.457	0.540	0.850	0.081	0.813	0.596	0.206	0.234	0.182	0.343	0.205	0.451
		95% CI	(-0.009 - 0.019)	(-0.109 - 0.184)	(-0.062 - 0.074)	(-0.012 - 0.151)	(-0.065 - 0.052)	(-0.014 - 0.023)	(-0.085 - 0.021)	(-0.178 - 0.05)	(-0.052 - 0.011)	(-0.018 - 0.047)	(-0.09 - 0.022)	(-0.124 - 0.06)
	C2		0.800	0.903	0.825	0.748	0.784	0.783	0.251	0.164	0.208	0.246	0.340	0.203
		P-value	0.647	0.874	0.620	0.139	0.871	0.736	0.779	0.596	0.343	0.343	0.319	0.787
		95% CI	(-0.01 - 0.016)	(-0.12 - 0.136)	(-0.037 - 0.058)	(-0.035 - 0.188)	(-0.043 - 0.049)	(-0.014 - 0.02)	(-0.072 - 0.056)	(-0.2 - 0.329)	(-0.029 - 0.075)	(-0.018 - 0.047)	(-0.106 - 0.038)	(-0.113 - 0.088)
	C3		0.786	0.833	0.824	0.774	0.774	0.768	0.260	0.329	0.216	0.246	0.344	0.195
		P-value	0.016	0.254	0.502	0.140	0.517	0.041	0.410	0.168	0.774	0.343	0.095	0.858
		95% CI	(0.004 - 0.03)	(-0.078 - 0.234)	(-0.026 - 0.049)	(-0.024 - 0.125)	(-0.032 - 0.058)	(0.001 - 0.034)	(-0.064 - 0.029)	(-0.251 - 0.051)	(-0.098 - 0.128)	(-0.018 - 0.047)	(-0.084 - 0.008)	(-0.06 - 0.051)
Respiratory SOFA score	R		0.900	0.633	0.941	0.806	0.858	0.903	0.685	0.360	0.706	0.503	0.653	0.687
	T		0.896	0.610	0.950	0.785	0.845	0.902	0.674	0.260	0.689	0.492	0.665	0.682
		P-value	0.079	0.530	0.279	0.432	0.080	0.847	0.142	0.343	0.343	0.343	0.673	0.576
		95% CI	(-0.0 - 0.007)	(-0.065 - 0.112)	(-0.027 - 0.009)	(-0.039 - 0.081)	(-0.002 - 0.028)	(-0.006 - 0.007)	(-0.004 - 0.025)	(-0.126 - 0.326)	(-0.021 - 0.054)	(-0.013 - 0.033)	(-0.074 - 0.05)	(-0.015 - 0.025)
	C1		0.899	0.654	0.953	0.792	0.850	0.901	0.673	0.293	0.710	0.492	0.661	0.672
		P-value	0.553	0.550	0.085	0.663	0.236	0.579	0.307	0.343	0.343	0.343	0.871	0.496
		95% CI	(-0.003 - 0.006)	(-0.107 - 0.064)	(-0.026 - 0.002)	(-0.056 - 0.082)	(-0.006 - 0.021)	(-0.006 - 0.01)	(-0.012 - 0.035)	(-0.084 - 0.218)	(-0.013 - 0.005)	(-0.013 - 0.033)	(-0.117 - 0.101)	(-0.034 - 0.065)
	C2		0.900	0.648	0.947	0.815	0.853	0.900	0.679	0.260	0.668	0.503	0.708	0.691
		P-value	0.951	0.612	0.473	0.518	0.679	0.698	0.691	0.343	0.343		0.224	0.821
		95% CI	(-0.007 - 0.007)	(-0.083 - 0.054)	(-0.023 - 0.012)	(-0.041 - 0.023)	(-0.023 - 0.034)	(-0.012 - 0.017)	(-0.025 - 0.036)	(-0.126 - 0.326)	(-0.047 - 0.122)	(0.0 - 0.0)	(-0.152 - 0.041)	(-0.034 - 0.028)
	C3		0.899	0.618	0.950	0.812	0.844	0.903	0.697	0.360	0.746	0.492	0.684	0.700
		P-value	0.669	0.571	0.175	0.698	0.306	0.983	0.064		0.395	0.343	0.501	0.179
		95% CI	(-0.002 - 0.004)	(-0.048 - 0.078)	(-0.022 - 0.005)	(-0.039 - 0.028)	(-0.016 - 0.045)	(-0.01 - 0.01)	(-0.026 - 0.001)	(0.0 - 0.0)	(-0.142 - 0.062)	(-0.013 - 0.033)	(-0.132 - 0.07)	(-0.033 - 0.007)
Overall SOFA score increase	R		0.698	0.787	0.661	0.767	0.680	0.685	0.108	0.070	0.142	0.075	0.047	0.111
	T		0.714	0.791	0.714	0.820	0.734	0.695	0.102	0.010	0.111	0.175	0.101	0.105
		P-value	0.167	0.944	0.438	0.468	0.058	0.509	0.648	0.343	0.169	0.343	0.224	0.642
		95% CI	(-0.039 - 0.008)	(-0.138 - 0.13)	(-0.199 - 0.094)	(-0.218 - 0.111)	(-0.111 - 0.002)	(-0.046 - 0.025)	(-0.021 - 0.032)	(-0.076 - 0.196)	(-0.016 - 0.078)	(-0.326 - 0.126)	(-0.148 - 0.04)	(-0.021 - 0.033)
	C1		0.693	0.780	0.672	0.772	0.727	0.673	0.096	0.010	0.146	0.183	0.054	0.095
		P-value	0.657	0.924	0.720	0.837	0.047	0.518	0.325	0.343	0.820	0.304	0.769	0.355
		95% CI	(-0.02 - 0.03)	(-0.158 - 0.172)	(-0.082 - 0.059)	(-0.066 - 0.055)	(-0.094 - -0.001)	(-0.028 - 0.052)	(-0.013 - 0.036)	(-0.076 - 0.196)	(-0.046 - 0.037)	(-0.333 - 0.117)	(-0.063 - 0.048)	(-0.021 - 0.053)
	C2		0.685	0.816	0.672	0.781	0.698	0.674	0.096	0.010	0.132	0.175	0.034	0.100
		P-value	0.071	0.832	0.844	0.891	0.473	0.289	0.293	0.343	0.681	0.343	0.343	0.455
		95% CI	(-0.001 - 0.028)	(-0.356 - 0.297)	(-0.131 - 0.109)	(-0.246 - 0.218)	(-0.072 - 0.036)	(-0.011 - 0.032)	(-0.012 - 0.035)	(-0.076 - 0.196)	(-0.043 - 0.063)	(-0.326 - 0.126)	(-0.016 - 0.041)	(-0.022 - 0.044)
	C3		0.699	0.765	0.673	0.755	0.699	0.690	0.102	0.010	0.139	0.175	0.074	0.108
		P-value	0.898	0.715	0.808	0.860	0.354	0.602	0.496	0.343	0.907	0.343	0.550	0.831
		95% CI	(-0.022 - 0.019)	(-0.115 - 0.157)	(-0.121 - 0.097)	(-0.142 - 0.166)	(-0.065 - 0.026)	(-0.027 - 0.017)	(-0.013 - 0.025)	(-0.076 - 0.196)	(-0.059 - 0.065)	(-0.326 - 0.126)	(-0.128 - 0.073)	(-0.023 - 0.028)

Table B.25: F1-score and accuracy in the test set with RF, in the Thermometry study.

Task	Feature		F1-score						Accuracy					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.341	0.244	0.264	0.305	0.421	0.269	0.785	0.793	0.825	0.798	0.738	0.788
	T		0.354	0.358	0.257	0.278	0.439	0.266	0.782	0.802	0.832	0.766	0.741	0.780
		P-value	0.536	0.257	0.793	0.177	0.782	0.928	0.632	0.667	0.776	0.170	0.829	0.439
		95% CI	(-0.056 - 0.031)	(-0.327 - 0.099)	(-0.057 - 0.073)	(-0.014 - 0.067)	(-0.155 - 0.12)	(-0.081 - 0.088)	(-0.011 - 0.017)	(-0.059 - 0.04)	(-0.06 - 0.046)	(-0.016 - 0.08)	(-0.039 - 0.032)	(-0.015 - 0.032)
	C1		0.368	0.327	0.276	0.282	0.451	0.294	0.785	0.812	0.838	0.751	0.753	0.787
		P-value	0.338	0.249	0.644	0.170	0.328	0.626	0.989	0.185	0.409	0.088	0.245	0.912
		95% CI	(-0.086 - 0.033)	(-0.233 - 0.069)	(-0.067 - 0.044)	(-0.012 - 0.058)	(-0.096 - 0.036)	(-0.133 - 0.085)	(-0.017 - 0.016)	(-0.051 - 0.012)	(-0.049 - 0.022)	(-0.009 - 0.103)	(-0.044 - 0.013)	(-0.028 - 0.031)
	C2		0.338	0.192	0.218	0.275	0.461	0.272	0.781	0.781	0.824	0.762	0.746	0.785
		P-value	0.913	0.688	0.188	0.186	0.354	0.959	0.667	0.649	0.954	0.178	0.606	0.840
		95% CI	(-0.067 - 0.074)	(-0.235 - 0.34)	(-0.027 - 0.121)	(-0.017 - 0.076)	(-0.132 - 0.052)	(-0.122 - 0.117)	(-0.016 - 0.024)	(-0.045 - 0.069)	(-0.048 - 0.05)	(-0.02 - 0.092)	(-0.043 - 0.026)	(-0.03 - 0.036)
	C3		0.344	0.361	0.219	0.294	0.459	0.256	0.777	0.816	0.803	0.784	0.751	0.772
		P-value	0.925	0.173	0.519	0.343	0.166	0.691	0.283	0.181	0.572	0.343	0.173	0.183
		95% CI	(-0.058 - 0.053)	(-0.295 - 0.062)	(-0.109 - 0.2)	(-0.013 - 0.035)	(-0.095 - 0.019)	(-0.06 - 0.086)	(-0.009 - 0.026)	(-0.061 - 0.014)	(-0.063 - 0.107)	(-0.018 - 0.047)	(-0.033 - 0.007)	(-0.009 - 0.042)
Respiratory SOFA score	R		0.746	0.342	0.730	0.527	0.709	0.743	0.850	0.848	0.835	0.687	0.800	0.863
	T		0.740	0.269	0.736	0.516	0.684	0.747	0.847	0.760	0.840	0.678	0.765	0.867
		P-value	0.396	0.508	0.727	0.343	0.390	0.710	0.481	0.405	0.712	0.343	0.099	0.451
		95% CI	(-0.009 - 0.021)	(-0.167 - 0.314)	(-0.041 - 0.03)	(-0.013 - 0.034)	(-0.038 - 0.088)	(-0.023 - 0.016)	(-0.006 - 0.013)	(-0.141 - 0.319)	(-0.033 - 0.023)	(-0.012 - 0.031)	(-0.008 - 0.078)	(-0.015 - 0.007)
	C1		0.736	0.274	0.748	0.496	0.674	0.740	0.844	0.767	0.849	0.667	0.766	0.863
		P-value	0.332	0.218	0.121	0.190	0.399	0.823	0.292	0.268	0.142	0.169	0.151	0.962
		95% CI	(-0.011 - 0.03)	(-0.048 - 0.184)	(-0.041 - 0.006)	(-0.018 - 0.079)	(-0.055 - 0.125)	(-0.033 - 0.04)	(-0.006 - 0.017)	(-0.075 - 0.239)	(-0.034 - 0.006)	(-0.011 - 0.052)	(-0.015 - 0.083)	(-0.014 - 0.013)
	C2		0.744	0.269	0.711	0.522	0.730	0.750	0.849	0.760	0.822	0.683	0.801	0.866
		P-value	0.856	0.508	0.504	0.343	0.478	0.485	0.863	0.405	0.477	0.343	0.944	0.447
		95% CI	(-0.02 - 0.024)	(-0.167 - 0.314)	(-0.042 - 0.079)	(-0.006 - 0.015)	(-0.082 - 0.042)	(-0.027 - 0.014)	(-0.011 - 0.012)	(-0.141 - 0.319)	(-0.026 - 0.052)	(-0.006 - 0.015)	(-0.029 - 0.027)	(-0.013 - 0.006)
	C3		0.758	0.331	0.770	0.516	0.715	0.767	0.857	0.837	0.854	0.678	0.787	0.876
		P-value	0.094	0.343	0.327	0.343	0.842	<.01	0.129	0.343	0.369	0.343	0.447	<.01
		95% CI	(-0.026 - 0.002)	(-0.014 - 0.037)	(-0.128 - 0.047)	(-0.013 - 0.034)	(-0.07 - 0.058)	(-0.038 - -0.009)	(-0.016 - 0.002)	(-0.014 - 0.036)	(-0.063 - 0.026)	(-0.012 - 0.031)	(-0.023 - 0.048)	(-0.022 - -0.004)
Overall SOFA score increase	R		0.183	0.093	0.179	0.086	0.083	0.185	0.754	0.729	0.764	0.678	0.723	0.766
	T		0.171	0.018	0.154	0.186	0.159	0.174	0.750	0.696	0.760	0.687	0.731	0.762
		P-value	0.557	0.343	0.172	0.343	0.271	0.613	0.432	0.343	0.570	0.343	0.328	0.510
		95% CI	(-0.031 - 0.054)	(-0.095 - 0.245)	(-0.013 - 0.063)	(-0.326 - 0.126)	(-0.222 - 0.07)	(-0.036 - 0.057)	(-0.007 - 0.015)	(-0.042 - 0.109)	(-0.011 - 0.019)	(-0.03 - 0.011)	(-0.024 - 0.009)	(-0.01 - 0.018)
	C1		0.164	0.018	0.189	0.191	0.092	0.163	0.751	0.696	0.772	0.692	0.716	0.763
		P-value	0.303	0.343	0.757	0.318	0.841	0.394	0.271	0.343	0.323	0.187	0.453	0.445
		95% CI	(-0.02 - 0.056)	(-0.095 - 0.245)	(-0.084 - 0.063)	(-0.33 - 0.12)	(-0.104 - 0.087)	(-0.033 - 0.075)	(-0.003 - 0.01)	(-0.042 - 0.109)	(-0.026 - 0.009)	(-0.036 - 0.008)	(-0.014 - 0.029)	(-0.007 - 0.015)
	C2		0.164	0.018	0.169	0.186	0.060	0.171	0.752	0.696	0.767	0.687	0.717	0.765
		P-value	0.285	0.343	0.800	0.343	0.315	0.549	0.323	0.343	0.722	0.343	0.181	0.681
		95% CI	(-0.018 - 0.054)	(-0.095 - 0.245)	(-0.074 - 0.093)	(-0.326 - 0.126)	(-0.026 - 0.073)	(-0.035 - 0.062)	(-0.003 - 0.008)	(-0.042 - 0.109)	(-0.022 - 0.016)	(-0.03 - 0.011)	(-0.003 - 0.015)	(-0.006 - 0.009)
	C3		0.173	0.018	0.155	0.186	0.122	0.183	0.753	0.696	0.760	0.687	0.725	0.766
		P-value	0.438	0.343	0.512	0.343	0.595	0.900	0.532	0.343	0.755	0.343	0.910	0.961
		95% CI	(-0.018 - 0.038)	(-0.095 - 0.245)	(-0.056 - 0.104)	(-0.326 - 0.126)	(-0.199 - 0.121)	(-0.031 - 0.035)	(-0.004 - 0.007)	(-0.042 - 0.109)	(-0.022 - 0.03)	(-0.03 - 0.011)	(-0.035 - 0.031)	(-0.004 - 0.005)

Table B.26: AUROC and recall in the test set with XGBoost, in the Thermometry study.

Task	Feature		AUROC						Recall					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.773	0.761	0.755	0.789	0.790	0.758	0.454	0.193	0.407	0.460	0.515	0.426
	T		0.767	0.778	0.715	0.822	0.794	0.747	0.446	0.381	0.355	0.332	0.470	0.423
		P-value	0.318	0.754	0.226	0.287	0.859	0.238	0.472	0.169	0.410	0.283	0.357	0.862
		95% CI	(-0.008 - 0.021)	(-0.15 - 0.116)	(-0.03 - 0.11)	(-0.102 - 0.037)	(-0.051 - 0.043)	(-0.009 - 0.031)	(-0.016 - 0.032)	(-0.473 - 0.096)	(-0.085 - 0.191)	(-0.126 - 0.383)	(-0.06 - 0.149)	(-0.038 - 0.045)
	C1		0.782	0.746	0.732	0.794	0.800	0.766	0.462	0.210	0.412	0.460	0.528	0.416
		P-value	0.137	0.761	0.448	0.738	0.588	0.409	0.592	0.343	0.796	1.000	0.684	0.747
		95% CI	(-0.021 - 0.003)	(-0.104 - 0.133)	(-0.044 - 0.091)	(-0.034 - 0.026)	(-0.045 - 0.027)	(-0.027 - 0.012)	(-0.041 - 0.025)	(-0.054 - 0.021)	(-0.04 - 0.031)	(-0.112 - 0.112)	(-0.084 - 0.057)	(-0.059 - 0.08)
	C2		0.781	0.799	0.732	0.798	0.820	0.759	0.448	0.426	0.402	0.360	0.474	0.397
		P-value	0.225	0.510	0.386	0.445	0.110	0.912	0.760	0.094	0.943	0.394	0.308	0.188
		95% CI	(-0.021 - 0.006)	(-0.177 - 0.1)	(-0.034 - 0.079)	(-0.033 - 0.017)	(-0.068 - 0.008)	(-0.017 - 0.015)	(-0.038 - 0.05)	(-0.515 - 0.049)	(-0.144 - 0.154)	(-0.153 - 0.353)	(-0.044 - 0.125)	(-0.017 - 0.075)
	C3		0.776	0.795	0.684	0.768	0.770	0.764	0.454	0.376	0.299	0.460	0.428	0.451
		P-value	0.542	0.389	0.167	0.528	0.266	0.347	0.996	0.170	0.169	1.000	0.040	0.274
		95% CI	(-0.013 - 0.007)	(-0.129 - 0.06)	(-0.036 - 0.177)	(-0.059 - 0.101)	(-0.019 - 0.061)	(-0.019 - 0.007)	(-0.044 - 0.044)	(-0.461 - 0.095)	(-0.056 - 0.272)	(-0.112 - 0.112)	(0.005 - 0.168)	(-0.072 - 0.023)
Respiratory SOFA score	R		0.888	0.634	0.892	0.788	0.841	0.892	0.746	0.323	0.755	0.553	0.717	0.767
	T		0.890	0.644	0.924	0.795	0.852	0.890	0.738	0.267	0.783	0.519	0.743	0.743
		P-value	0.599	0.810	0.244	0.199	0.578	0.380	0.599	0.635	0.334	0.392	0.502	0.331
		95% CI	(-0.007 - 0.004)	(-0.11 - 0.09)	(-0.089 - 0.026)	(-0.018 - 0.005)	(-0.056 - 0.033)	(-0.003 - 0.008)	(-0.025 - 0.041)	(-0.204 - 0.317)	(-0.09 - 0.034)	(-0.051 - 0.119)	(-0.109 - 0.058)	(-0.028 - 0.076)
	C1		0.889	0.656	0.924	0.785	0.844	0.892	0.743	0.343	0.803	0.503	0.738	0.764
		P-value	0.881	0.660	0.297	0.728	0.786	0.940	0.869	0.343	0.567	0.485	0.277	0.782
		95% CI	(-0.008 - 0.007)	(-0.143 - 0.099)	(-0.096 - 0.033)	(-0.015 - 0.021)	(-0.032 - 0.025)	(-0.009 - 0.008)	(-0.028 - 0.033)	(-0.065 - 0.025)	(-0.228 - 0.133)	(-0.106 - 0.207)	(-0.061 - 0.02)	(-0.02 - 0.026)
	C2		0.888	0.686	0.914	0.783	0.840	0.890	0.747	0.343	0.790	0.503	0.684	0.767
		P-value	0.913	0.387	0.369	0.457	0.938	0.552	0.864	0.343	0.388	0.485	0.506	0.995
		95% CI	(-0.006 - 0.006)	(-0.193 - 0.089)	(-0.074 - 0.03)	(-0.01 - 0.021)	(-0.019 - 0.02)	(-0.005 - 0.009)	(-0.017 - 0.014)	(-0.065 - 0.025)	(-0.123 - 0.052)	(-0.106 - 0.207)	(-0.076 - 0.144)	(-0.035 - 0.035)
	C3		0.885	0.590	0.931	0.701	0.837	0.888	0.733	0.310	0.805	0.492	0.729	0.742
		P-value	0.198	0.571	0.238	0.194	0.689	0.246	0.446	0.751	0.343	0.395	0.737	0.470
		95% CI	(-0.002 - 0.01)	(-0.142 - 0.23)	(-0.109 - 0.031)	(-0.056 - 0.229)	(-0.015 - 0.021)	(-0.003 - 0.01)	(-0.023 - 0.048)	(-0.079 - 0.105)	(-0.163 - 0.063)	(-0.093 - 0.214)	(-0.092 - 0.068)	(-0.049 - 0.099)
Overall SOFA score increase	R		0.670	0.749	0.650	0.759	0.701	0.665	0.338	0.250	0.284	0.246	0.302	0.338
	T		0.676	0.842	0.517	0.713	0.726	0.677	0.359	0.321	0.207	0.296	0.338	0.362
		P-value	0.443	0.469	0.041	0.595	0.516	0.257	0.351	0.611	0.177	0.343	0.548	0.456
		95% CI	(-0.026 - 0.012)	(-0.388 - 0.202)	(0.007 - 0.26)	(-0.151 - 0.244)	(-0.112 - 0.06)	(-0.035 - 0.011)	(-0.07 - 0.027)	(-0.376 - 0.234)	(-0.042 - 0.195)	(-0.163 - 0.063)	(-0.167 - 0.095)	(-0.094 - 0.046)
	C1		0.666	0.739	0.555	0.787	0.636	0.668	0.358	0.281	0.241	0.426	0.260	0.358
		P-value	0.709	0.951	0.053	0.707	0.031	0.827	0.593	0.806	0.380	0.123	0.269	0.660
		95% CI	(-0.02 - 0.028)	(-0.347 - 0.365)	(-0.002 - 0.192)	(-0.198 - 0.142)	(0.007 - 0.123)	(-0.033 - 0.027)	(-0.1 - 0.06)	(-0.309 - 0.247)	(-0.061 - 0.146)	(-0.418 - 0.059)	(-0.039 - 0.123)	(-0.117 - 0.078)
	C2		0.682	0.805	0.570	0.814	0.689	0.683	0.363	0.201	0.265	0.309	0.302	0.360
		P-value	0.189	0.697	0.355	0.198	0.607	0.074	0.385	0.513	0.720	0.217	1.000	0.511
		95% CI	(-0.033 - 0.007)	(-0.397 - 0.283)	(-0.106 - 0.268)	(-0.145 - 0.036)	(-0.037 - 0.06)	(-0.038 - 0.002)	(-0.088 - 0.037)	(-0.113 - 0.211)	(-0.096 - 0.134)	(-0.17 - 0.044)	(-0.031 - 0.031)	(-0.096 - 0.051)
	C3		0.673	0.793	0.646	0.814	0.699	0.660	0.329	0.277	0.366	0.322	0.215	0.330
		P-value	0.752	0.745	0.967	0.563	0.960	0.625	0.789	0.756	0.501	0.230	0.083	0.834
		95% CI	(-0.026 - 0.02)	(-0.365 - 0.276)	(-0.211 - 0.219)	(-0.27 - 0.16)	(-0.068 - 0.072)	(-0.016 - 0.025)	(-0.065 - 0.083)	(-0.215 - 0.162)	(-0.347 - 0.183)	(-0.208 - 0.057)	(-0.014 - 0.188)	(-0.081 - 0.099)

Table B.27: F1-score and accuracy in the test set with XGBoost, in the Thermometry study.

Task	Feature		F1-score						Accuracy					
			All	A	B	HL	O	W	All	A	B	HL	O	W
In-hospital mortality	R		0.473	0.210	0.363	0.431	0.590	0.426	0.763	0.706	0.771	0.710	0.754	
	T		0.478	0.344	0.335	0.326	0.526	0.437	0.772	0.735	0.804	0.696	0.739	
		P-value	0.413	0.235	0.615	0.412	0.217	0.408	0.034	0.501	0.347	0.730	0.553	
		95% CI	(-0.02 - 0.009)	(-0.373 - 0.105)	(-0.093 - 0.148)	(-0.172 - 0.384)	(-0.044 - 0.171)	(-0.042 - 0.019)	(-0.016 - -0.001)	(-0.12 - 0.064)	(-0.108 - 0.042)	(-0.077 - 0.106)	(-0.038 - 0.066)	
	C1		0.488	0.213	0.376	0.468	0.578	0.432	0.772	0.691	0.788	0.762	0.758	
		P-value	0.325	0.933	0.597	0.617	0.741	0.810	0.317	0.554	0.559	0.117	0.851	
		95% CI	(-0.049 - 0.018)	(-0.076 - 0.07)	(-0.066 - 0.04)	(-0.195 - 0.122)	(-0.067 - 0.09)	(-0.059 - 0.047)	(-0.03 - 0.011)	(-0.043 - 0.075)	(-0.078 - 0.045)	(-0.121 - 0.016)	(-0.056 - 0.047)	
	C2		0.471	0.362	0.385	0.355	0.544	0.409	0.765	0.776	0.760	0.702	0.753	
		P-value	0.919	0.191	0.743	0.549	0.302	0.312	0.774	0.328	0.424	0.832	0.982	
		95% CI	(-0.037 - 0.041)	(-0.396 - 0.091)	(-0.165 - 0.122)	(-0.202 - 0.355)	(-0.048 - 0.139)	(-0.018 - 0.051)	(-0.02 - 0.015)	(-0.225 - 0.085)	(-0.019 - 0.041)	(-0.078 - 0.095)	(-0.039 - 0.04)	
	C3		0.476	0.362	0.297	0.462	0.468	0.451	0.767	0.777	0.771	0.740	0.705	
		P-value	0.859	0.191	0.316	0.671	0.050	0.216	0.713	0.303	1.000	0.496	0.049	
		95% CI	(-0.048 - 0.041)	(-0.396 - 0.091)	(-0.075 - 0.207)	(-0.19 - 0.128)	(0.0 - 0.244)	(-0.068 - 0.018)	(-0.026 - 0.018)	(-0.219 - 0.077)	(-0.09 - 0.09)	(-0.126 - 0.066)	(0.0 - 0.096)	
Respiratory SOFA score	R		0.755	0.303	0.767	0.562	0.686	0.768	0.842	0.794	0.857	0.696	0.746	
	T		0.761	0.256	0.802	0.532	0.730	0.757	0.850	0.727	0.878	0.692	0.797	
		P-value	0.595	0.695	0.163	0.443	0.227	0.512	0.268	0.543	0.160	0.867	0.071	
		95% CI	(-0.03 - 0.018)	(-0.216 - 0.31)	(-0.089 - 0.017)	(-0.054 - 0.114)	(-0.121 - 0.033)	(-0.025 - 0.047)	(-0.023 - 0.007)	(-0.174 - 0.309)	(-0.053 - 0.01)	(-0.051 - 0.059)	(-0.109 - 0.006)	
	C1		0.763	0.316	0.786	0.514	0.723	0.775	0.850	0.797	0.855	0.671	0.783	
		P-value	0.484	0.448	0.683	0.440	0.052	0.101	0.327	0.798	0.953	0.522	0.057	
		95% CI	(-0.034 - 0.017)	(-0.049 - 0.024)	(-0.119 - 0.082)	(-0.086 - 0.181)	(-0.075 - 0.0)	(-0.017 - 0.002)	(-0.024 - 0.009)	(-0.027 - 0.021)	(-0.062 - 0.065)	(-0.059 - 0.108)	(-0.076 - 0.001)	
	C2		0.765	0.318	0.801	0.514	0.674	0.778	0.850	0.797	0.878	0.673	0.750	
		P-value	0.258	0.343	0.248	0.430	0.793	0.242	0.243	0.773	0.285	0.537	0.891	
		95% CI	(-0.028 - 0.008)	(-0.051 - 0.02)	(-0.096 - 0.028)	(-0.084 - 0.18)	(-0.082 - 0.104)	(-0.029 - 0.008)	(-0.022 - 0.006)	(-0.027 - 0.02)	(-0.062 - 0.02)	(-0.057 - 0.102)	(-0.075 - 0.066)	
	C3		0.750	0.310	0.781	0.508	0.689	0.757	0.842	0.811	0.853	0.668	0.739	
		P-value	0.711	0.832	0.517	0.373	0.909	0.630	0.952	0.661	0.686	0.447	0.798	
		95% CI	(-0.023 - 0.032)	(-0.083 - 0.068)	(-0.06 - 0.033)	(-0.076 - 0.184)	(-0.063 - 0.056)	(-0.037 - 0.058)	(-0.014 - 0.015)	(-0.1 - 0.067)	(-0.019 - 0.027)	(-0.051 - 0.105)	(-0.049 - 0.062)	
Overall SOFA score increase	R		0.387	0.321	0.230	0.230	0.355	0.380	0.725	0.762	0.647	0.653	0.727	
	T		0.404	0.336	0.214	0.324	0.407	0.394	0.723	0.799	0.678	0.742	0.738	
		P-value	0.418	0.914	0.735	0.227	0.415	0.614	0.833	0.527	0.586	0.066	0.658	
		95% CI	(-0.063 - 0.029)	(-0.337 - 0.305)	(-0.089 - 0.122)	(-0.258 - 0.07)	(-0.191 - 0.086)	(-0.072 - 0.045)	(-0.021 - 0.025)	(-0.163 - 0.09)	(-0.154 - 0.092)	(-0.186 - 0.007)	(-0.065 - 0.043)	
	C1		0.398	0.294	0.248	0.426	0.293	0.387	0.724	0.687	0.657	0.779	0.698	
		P-value	0.756	0.838	0.623	0.057	0.179	0.851	0.918	0.244	0.857	0.029	0.295	
		95% CI	(-0.085 - 0.064)	(-0.256 - 0.309)	(-0.095 - 0.06)	(-0.399 - 0.007)	(-0.034 - 0.159)	(-0.093 - 0.078)	(-0.027 - 0.029)	(-0.061 - 0.211)	(-0.136 - 0.116)	(-0.236 - -0.016)	(-0.03 - 0.087)	
	C2		0.406	0.224	0.234	0.321	0.334	0.402	0.729	0.745	0.622	0.727	0.725	
		P-value	0.365	0.282	0.942	0.082	0.265	0.356	0.601	0.633	0.529	0.049	0.915	
		95% CI	(-0.064 - 0.026)	(-0.095 - 0.288)	(-0.11 - 0.103)	(-0.196 - 0.014)	(-0.019 - 0.062)	(-0.075 - 0.03)	(-0.019 - 0.012)	(-0.061 - 0.095)	(-0.061 - 0.111)	(-0.148 - -0.001)	(-0.035 - 0.039)	
	C3		0.382	0.322	0.346	0.339	0.272	0.370	0.727	0.708	0.713	0.751	0.721	
		95% CI	0.863	0.992	0.303	0.208	0.101	0.786	0.895	0.531	0.199	0.072	0.732	
		P-value	(-0.064 - 0.074)	(-0.244 - 0.242)	(-0.356 - 0.124)	(-0.291 - 0.073)	(-0.02 - 0.186)	(-0.07 - 0.09)	(-0.025 - 0.022)	(-0.133 - 0.241)	(-0.174 - 0.042)	(-0.207 - 0.011)	(-0.035 - 0.048)	

Table B.28: Performance metrics across disparity groups in the test set with LR, in the Thermometry study.

Task	Feature		AUROC				Recall				F1-score				Accuracy			
			(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)
In-hospital mortality	R		0.787	0.811	0.775	0.860	0.600	0.726	0.664	0.785	0.496	0.582	0.481	0.615	0.751	0.749	0.693	0.721
	T		0.798	0.815	0.778	0.865	0.633	0.736	0.681	0.742	0.512	0.581	0.499	0.629	0.745	0.745	0.706	0.760
		P-value	0.389	0.312	0.495	0.419	0.343	0.178	0.343	0.238	0.343	0.884	0.099	0.521	0.685	0.377	0.071	0.012
		95% CI	(-0.04 - 0.017)	(-0.01 - 0.004)	(-0.01 - 0.005)	(- 0.019 - 0.009)	(- 0.109 - 0.042)	(- 0.027 - 0.006)	(- 0.058 - 0.022)	(- 0.034 - 0.119)	(- 0.054 - 0.021)	(- 0.011 - 0.013)	(-0.04 - 0.004)	(- 0.064 - 0.035)	(- 0.025 - 0.036)	(- 0.007 - 0.016)	(- 0.027 - 0.001)	(- 0.066 - - 0.011)
	C1		0.819	0.775	0.804	0.770	0.683	0.718	0.715	0.508	0.559	0.527	0.551	0.354	0.672	0.707	0.737	0.605
		P-value	0.784	0.223	0.097	0.208	0.533	0.889	0.281	0.057	0.652	0.227	0.030	0.023	0.371	0.159	0.020	0.026
		95% CI	(- 0.289 - 0.225)	(- 0.027 - 0.1)	(- 0.063 - 0.006)	(- 0.065 - 0.248)	(- 0.374 - 0.207)	(- 0.114 - 0.129)	(- 0.154 - 0.05)	(- 0.011 - 0.565)	(- 0.369 - 0.243)	(- 0.041 - 0.151)	(- 0.131 - - 0.008)	(0.045 - 0.475)	(- 0.111 - 0.27)	(-0.02 - 0.104)	(-0.08 - - 0.009)	(0.017 - 0.216)
	C2		0.823	0.798	0.804	0.755	0.683	0.713	0.719	0.474	0.561	0.537	0.556	0.339	0.699	0.738	0.726	0.629
		P-value	0.734	0.543	0.115	0.198	0.547	0.727	0.226	0.041	0.612	0.176	0.020	0.014	0.485	0.614	0.069	0.191
		95% CI	(- 0.269 - 0.197)	(- 0.034 - 0.06)	(- 0.066 - 0.009)	(- 0.074 - 0.288)	(- 0.385 - 0.218)	(- 0.065 - 0.09)	(- 0.151 - 0.041)	(0.015 - 0.606)	(- 0.345 - 0.215)	(- 0.024 - 0.114)	(- 0.135 - - 0.015)	(0.07 - 0.482)	(-0.11 - 0.214)	(- 0.039 - 0.062)	(- 0.068 - 0.003)	(- 0.055 - 0.239)
	C3		0.793	0.811	0.805	0.678	0.645	0.755	0.708	0.771	0.526	0.541	0.554	0.588	0.673	0.745	0.722	0.660
		P-value	0.850	0.993	0.100	0.060	0.782	0.528	0.310	0.898	0.810	0.243	0.034	0.828	0.206	0.792	0.047	0.462
		95% CI	(- 0.268 - 0.226)	(-0.05 - 0.051)	(- 0.067 - 0.007)	(-0.01 - 0.386)	(- 0.404 - 0.313)	(- 0.131 - 0.072)	(-0.14 - 0.05)	(- 0.222 - 0.25)	(- 0.312 - 0.251)	(- 0.033 - 0.116)	(-0.14 - - 0.007)	(- 0.247 - 0.302)	(- 0.051 - 0.207)	(- 0.031 - 0.039)	(- 0.058 - -0.0)	(-0.12 - 0.244)
Respiratory SOFA score	R		0.804	0.873	0.854	0.779	0.605	0.789	0.729	0.684	0.587	0.779	0.685	0.668	0.815	0.839	0.817	0.765
	T		0.833	0.873	0.849	0.764	0.751	0.797	0.687	0.571	0.663	0.776	0.663	0.600	0.797	0.834	0.811	0.742
		P-value	0.425	0.976	0.288	0.430	0.182	0.519	0.049	0.030	0.309	0.510	0.179	0.177	0.546	0.298	0.435	0.498
		95% CI	(- 0.105 - 0.048)	(- 0.008 - 0.007)	(- 0.005 - 0.014)	(- 0.026 - 0.056)	(- 0.374 - 0.082)	(- 0.032 - 0.018)	(0.0 - 0.084)	(0.014 - 0.212)	(- 0.235 - 0.083)	(- 0.008 - 0.015)	(- 0.012 - 0.056)	(- 0.037 - 0.173)	(- 0.048 - 0.085)	(- 0.005 - 0.014)	(-0.01 - 0.022)	(-0.05 - 0.095)
	C1		0.905	0.886	0.860	0.646	0.871	0.773	0.730	0.600	0.783	0.775	0.712	0.570	0.810	0.829	0.815	0.813
		P-value	0.274	0.393	0.743	0.200	0.121	0.572	0.983	0.569	0.237	0.882	0.354	0.483	0.921	0.726	0.881	0.457
		95% CI	(- 0.253 - 0.082)	(- 0.048 - 0.021)	(- 0.047 - 0.035)	(- 0.094 - 0.374)	(- 0.616 - 0.086)	(- 0.047 - 0.08)	(- 0.062 - 0.061)	(- 0.238 - 0.406)	(- 0.546 - 0.154)	(- 0.061 - 0.07)	(-0.09 - 0.036)	(- 0.206 - 0.403)	(- 0.118 - 0.129)	(- 0.052 - 0.072)	(- 0.026 - 0.03)	(-0.19 - 0.093)
	C2		0.896	0.853	0.867	0.593	0.817	0.738	0.737	0.483	0.748	0.740	0.711	0.472	0.818	0.827	0.811	0.794
		P-value	0.406	0.052	0.485	0.085	0.286	0.056	0.747	0.163	0.408	0.067	0.351	0.179	0.966	0.362	0.672	0.683
		95% CI	(- 0.278 - 0.125)	(-0.0 - 0.041)	(- 0.052 - 0.027)	(- 0.036 - 0.422)	(- 0.633 - 0.21)	(- 0.002 - 0.104)	(- 0.061 - 0.045)	(- 0.098 - 0.5)	(- 0.578 - 0.258)	(- 0.003 - 0.082)	(- 0.086 - 0.034)	(- 0.108 - 0.499)	(- 0.168 - 0.162)	(- 0.016 - 0.04)	(- 0.023 - 0.035)	(- 0.187 - 0.128)
	C3		0.843	0.854	0.873	0.728	0.745	0.721	0.753	0.742	0.772	0.680	0.750	0.620	0.821	0.797	0.838	0.726
		P-value	0.682	0.208	0.270	0.472	0.444	0.104	0.446	0.463	0.297	0.021	0.056	0.486	0.934	0.093	0.145	0.525
		95% CI	(- 0.248 - 0.172)	(- 0.013 - 0.05)	(- 0.056 - 0.018)	(- 0.103 - 0.205)	(- 0.534 - 0.255)	(- 0.017 - 0.153)	(- 0.091 - 0.043)	(- 0.228 - 0.113)	(- 0.561 - 0.193)	(-0.019 - 0.18)	(- 0.132 - 0.002)	(- 0.101 - 0.196)	(- 0.151 - 0.14)	(- 0.009 - 0.092)	(- 0.052 - 0.009)	(- 0.094 - 0.171)
Overall SOFA score increase	R		0.748	0.677	0.710	0.634	0.587	0.582	0.628	0.541	0.321	0.454	0.477	0.501	0.563	0.634	0.672	0.633
	T		0.744	0.677	0.714	0.628	0.637	0.559	0.673	0.565	0.399	0.439	0.492	0.510	0.637	0.628	0.669	0.627
		P-value	0.362	0.851	0.515	0.222	0.343	0.167	0.167	0.179	0.138	0.201	0.410	0.604	0.016	0.271	0.746	0.679
		95% CI	(- 0.007 - 0.016)	(- 0.006 - 0.005)	(- 0.015 - 0.008)	(- 0.004 - 0.016)	(- 0.163 - 0.063)	(- 0.012 - 0.058)	(- 0.113 - 0.023)	(-0.06 - 0.013)	(- 0.187 - 0.031)	(- 0.009 - 0.038)	(- 0.053 - 0.024)	(- 0.046 - 0.028)	(-0.13 - - 0.018)	(- 0.006 - 0.019)	(- 0.018 - 0.024)	(- 0.022 - 0.033)
	C1		0.658	0.685	0.685	0.708	0.250	0.516	0.609	0.447	0.183	0.396	0.474	0.379	0.690	0.637	0.641	0.668
		P-value	0.867	0.870	0.422	0.861	0.051	0.526	0.697	0.527	0.201	0.463	0.947	0.360	0.015	0.942	0.203	0.730
		95% CI	(- 0.237 - 0.274)	(- 0.122 - 0.105)	(- 0.043 - 0.094)	(- 0.239 - 0.276)	(- 0.001 - 0.675)	(- 0.161 - 0.293)	(- 0.086 - 0.124)	(- 0.229 - 0.417)	(- 0.088 - 0.362)	(- 0.112 - 0.227)	(- 0.089 - 0.095)	(- 0.164 - 0.407)	(- 0.222 - 0.031)	(- 0.078 - 0.073)	(-0.02 - 0.082)	(- 0.261 - 0.19)
	C2		0.686	0.669	0.710	0.618	0.350	0.546	0.651	0.443	0.250	0.443	0.473	0.389	0.678	0.638	0.629	0.749
		P-value	0.652	0.771	1.000	0.587	0.108	0.406	0.666	0.507	0.525	0.805	0.931	0.408	0.022	0.870	0.122	0.119
		95% CI	(- 0.125 - 0.187)	(- 0.048 - 0.062)	(- 0.061 - 0.061)	(- 0.267 - 0.411)	(- 0.063 - 0.537)	(- 0.058 - 0.132)	(- 0.139 - 0.093)	(- 0.223 - 0.419)	(- 0.171 - 0.313)	(- 0.082 - 0.103)	(- 0.095 - 0.102)	(- 0.179 - 0.403)	(-0.21 - - 0.021)	(- 0.058 - 0.05)	(- 0.014 - 0.1)	(-0.27 - 0.037)
	C3		0.628	0.686	0.708	0.653	0.275	0.540	0.632	0.429	0.257	0.438	0.490	0.362	0.679	0.657	0.635	0.725
		P-value	0.549	0.707	0.919	0.682	0.034	0.230	0.927	0.444	0.587	0.632	0.682	0.256	0.187	0.306	0.080	0.235
		95% CI	(- 0.194 - 0.331)	(- 0.061 - 0.043)	(- 0.056 - 0.061)	(- 0.156 - 0.223)	(0.029 - 0.594)	(- 0.032 - 0.116)	(- 0.096 - 0.088)	(- 0.206 - 0.432)	(- 0.193 - 0.322)	(- 0.058 - 0.09)	(- 0.085 - 0.058)	(-0.12 - 0.397)	(- 0.298 - 0.067)	(- 0.068 - 0.024)	(- 0.005 - 0.078)	(- 0.256 - 0.072)

Table B.29: Performance metrics across disparity groups in the test set with RF, in the Thermometry study.

Task	Feature		AUROC				Recall				F1-score				Accuracy			
			(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)
In-hospital mortality	R		0.773	0.814	0.785	0.832	0.140	0.221	0.253	0.285	0.183	0.318	0.333	0.371	0.768	0.784	0.793	0.734
	T		0.822	0.804	0.754	0.833	0.348	0.277	0.249	0.271	0.321	0.357	0.323	0.351	0.757	0.785	0.786	0.741
		P-value	0.362	0.445	0.011	0.967	0.168	0.119	0.821	0.604	0.278	0.293	0.717	0.502	0.768	0.830	0.635	0.639
		95% CI	(-0.162 - 0.065)	(-0.017 - 0.036)	(0.009 - 0.053)	(-0.038 - 0.036)	(-0.523 - 0.106)	(-0.129 - 0.017)	(-0.03 - 0.037)	(-0.043 - 0.07)	(-0.41 - 0.133)	(-0.116 - 0.039)	(-0.05 - 0.07)	(-0.045 - 0.086)	(-0.071 - 0.094)	(-0.017 - 0.014)	(-0.027 - 0.042)	(-0.043 - 0.028)
	C1		0.742	0.769	0.807	0.816	0.125	0.284	0.266	0.290	0.133	0.345	0.363	0.302	0.663	0.783	0.793	0.767
		P-value	0.764	0.157	0.098	0.986	0.905	0.454	0.695	0.975	0.739	0.774	0.462	0.689	0.131	0.978	0.965	0.698
		95% CI	(-0.196 - 0.258)	(-0.021 - 0.11)	(-0.048 - 0.005)	(-0.223 - 0.22)	(-0.262 - 0.292)	(-0.243 - 0.118)	(-0.086 - 0.06)	(-0.379 - 0.368)	(-0.276 - 0.375)	(-0.233 - 0.179)	(-0.115 - 0.057)	(-0.306 - 0.443)	(-0.038 - 0.249)	(-0.064 - 0.066)	(-0.035 - 0.036)	(-0.22 - 0.154)
	C2		0.784	0.813	0.790	0.824	0.142	0.299	0.202	0.240	0.185	0.387	0.283	0.271	0.701	0.810	0.762	0.797
		P-value	0.905	0.975	0.796	0.940	0.990	0.155	0.131	0.751	0.987	0.255	0.174	0.477	0.296	0.110	0.088	0.474
		95% CI	(-0.208 - 0.187)	(-0.048 - 0.05)	(-0.048 - 0.038)	(-0.214 - 0.228)	(-0.275 - 0.272)	(-0.192 - 0.036)	(-0.018 - 0.119)	(-0.264 - 0.354)	(-0.353 - 0.348)	(-0.197 - 0.059)	(-0.027 - 0.127)	(-0.205 - 0.405)	(-0.07 - 0.205)	(-0.06 - 0.007)	(-0.006 - 0.068)	(-0.256 - 0.129)
	C3		0.702	0.802	0.797	0.742	0.173	0.310	0.231	0.310	0.155	0.375	0.328	0.282	0.638	0.805	0.768	0.548
		P-value	0.630	0.496	0.333	0.337	0.808	0.031	0.695	0.873	0.794	0.214	0.936	0.542	0.042	0.257	0.233	0.075
		95% CI	(-0.212 - 0.33)	(-0.026 - 0.049)	(-0.036 - 0.014)	(-0.111 - 0.287)	(-0.327 - 0.262)	(-0.168 - -0.01)	(-0.098 - 0.14)	(-0.371 - 0.32)	(-0.206 - 0.261)	(-0.152 - 0.039)	(-0.145 - 0.156)	(-0.229 - 0.407)	(-0.006 - 0.255)	(-0.063 - 0.019)	(-0.019 - 0.069)	(-0.023 - 0.393)
Respiratory SOFA score	R		0.891	0.911	0.883	0.828	0.710	0.707	0.655	0.608	0.734	0.771	0.713	0.640	0.876	0.849	0.859	0.778
	T		0.903	0.906	0.881	0.835	0.657	0.716	0.641	0.526	0.690	0.778	0.704	0.565	0.860	0.853	0.857	0.731
		P-value	0.601	0.151	0.684	0.671	0.182	0.370	0.305	0.217	0.170	0.368	0.414	0.204	0.179	0.455	0.672	0.138
		95% CI	(-0.058 - 0.036)	(-0.002 - 0.012)	(-0.008 - 0.011)	(-0.046 - 0.031)	(-0.03 - 0.137)	(-0.03 - 0.012)	(-0.015 - 0.042)	(-0.058 - 0.221)	(-0.023 - 0.11)	(-0.024 - 0.01)	(-0.014 - 0.032)	(-0.049 - 0.2)	(-0.008 - 0.038)	(-0.016 - 0.008)	(-0.01 - 0.014)	(-0.018 - 0.113)
	C1		0.954	0.923	0.891	0.748	0.498	0.679	0.676	0.508	0.519	0.773	0.728	0.516	0.764	0.855	0.844	0.794
		P-value	0.240	0.457	0.522	0.399	0.331	0.444	0.397	0.434	0.303	0.963	0.461	0.346	0.122	0.749	0.095	0.693
		95% CI	(-0.14 - 0.04)	(-0.045 - 0.022)	(-0.036 - 0.019)	(-0.16 - 0.356)	(-0.255 - 0.678)	(-0.051 - 0.107)	(-0.076 - 0.033)	(-0.175 - 0.374)	(-0.229 - 0.658)	(-0.073 - 0.07)	(-0.059 - 0.029)	(-0.159 - 0.408)	(-0.036 - 0.258)	(-0.051 - 0.038)	(-0.003 - 0.035)	(-0.104 - 0.072)
	C2		0.948	0.888	0.908	0.785	0.717	0.670	0.679	0.417	0.717	0.747	0.730	0.437	0.845	0.850	0.846	0.806
		P-value	0.319	0.017	0.132	0.381	0.975	0.111	0.438	0.130	0.931	0.294	0.509	0.169	0.628	0.918	0.351	0.581
		95% CI	(-0.139 - 0.051)	(0.005 - 0.042)	(-0.06 - 0.009)	(-0.087 - 0.197)	(-0.459 - 0.446)	(-0.01 - 0.085)	(-0.093 - 0.044)	(-0.068 - 0.451)	(-0.394 - 0.426)	(-0.025 - 0.073)	(-0.072 - 0.038)	(-0.104 - 0.511)	(-0.108 - 0.17)	(-0.034 - 0.031)	(-0.017 - 0.042)	(-0.138 - 0.082)
	C3		0.942	0.896	0.907	0.792	0.741	0.685	0.723	0.617	0.800	0.748	0.776	0.589	0.833	0.861	0.867	0.740
		P-value	0.539	0.221	0.083	0.605	0.825	0.528	0.081	0.889	0.570	0.492	0.039	0.470	0.416	0.582	0.621	0.462
		95% CI	(-0.135 - 0.077)	(-0.011 - 0.041)	(-0.052 - 0.004)	(-0.115 - 0.187)	(-0.331 - 0.27)	(-0.054 - 0.098)	(-0.147 - 0.01)	(-0.151 - 0.133)	(-0.321 - 0.188)	(-0.05 - 0.096)	(-0.121 - 0.004)	(-0.102 - 0.204)	(-0.07 - 0.154)	(-0.063 - 0.038)	(-0.042 - 0.027)	(-0.075 - 0.152)
Overall SOFA score increase	R		0.764	0.673	0.696	0.649	0.253	0.107	0.088	0.144	0.250	0.174	0.151	0.211	0.760	0.749	0.774	0.693
	T		0.802	0.690	0.713	0.650	0.158	0.121	0.068	0.127	0.157	0.193	0.118	0.183	0.746	0.756	0.763	0.683
		P-value	0.449	0.291	0.346	0.982	0.394	0.479	0.141	0.557	0.437	0.499	0.139	0.519	0.329	0.340	0.025	0.180
		95% CI	(-0.143 - 0.069)	(-0.051 - 0.017)	(-0.055 - 0.022)	(-0.121 - 0.119)	(-0.145 - 0.335)	(-0.056 - 0.028)	(-0.008 - 0.049)	(-0.045 - 0.078)	(-0.166 - 0.353)	(-0.08 - 0.042)	(-0.013 - 0.078)	(-0.068 - 0.125)	(-0.017 - 0.047)	(-0.023 - 0.009)	(0.002 - 0.02)	(-0.006 - 0.026)
	C1		0.621	0.679	0.693	0.814	0.000	0.092	0.097	0.046	0.000	0.135	0.167	0.072	0.783	0.763	0.747	0.739
		P-value	0.245	0.875	0.921	0.173	0.082	0.816	0.723	0.066	0.057	0.660	0.678	0.081	0.730	0.542	0.051	0.614
		95% CI	(-0.125 - 0.405)	(-0.084 - 0.072)	(-0.053 - 0.058)	(-0.527 - 0.124)	(-0.04 - 0.546)	(-0.124 - 0.153)	(-0.064 - 0.046)	(-0.008 - 0.204)	(-0.009 - 0.509)	(-0.155 - 0.233)	(-0.102 - 0.069)	(-0.021 - 0.299)	(-0.166 - 0.121)	(-0.061 - 0.034)	(-0.0.0 - 0.056)	(-0.245 - 0.153)
	C2		0.791	0.677	0.688	0.773	0.000	0.106	0.100	0.000	0.000	0.176	0.172	0.000	0.753	0.744	0.759	0.680
		P-value	0.998	0.907	0.771	0.293	0.082	0.967	0.684	0.044	0.057	0.957	0.632	0.032	0.931	0.780	0.302	0.916
		95% CI	(-0.203 - 0.202)	(-0.063 - 0.057)	(-0.054 - 0.07)	(-0.755 - 0.297)	(-0.04 - 0.546)	(-0.047 - 0.049)	(-0.073 - 0.05)	(0.005 - 0.283)	(-0.009 - 0.509)	(-0.075 - 0.072)	(-0.117 - 0.075)	(0.023 - 0.4)	(-0.179 - 0.194)	(-0.034 - 0.044)	(-0.016 - 0.046)	(-0.262 - 0.289)
	C3		0.665	0.679	0.713	0.763	0.000	0.104	0.119	0.000	0.000	0.176	0.193	0.000	0.777	0.765	0.742	0.713
		P-value	0.086	0.884	0.408	0.473	0.082	0.880	0.373	0.044	0.057	0.957	0.406	0.032	0.848	0.290	0.097	0.827
		95% CI	(-0.022 - 0.246)	(-0.093 - 0.082)	(-0.062 - 0.028)	(-0.525 - 0.275)	(-0.04 - 0.546)	(-0.042 - 0.048)	(-0.104 - 0.043)	(0.005 - 0.283)	(-0.009 - 0.509)	(-0.075 - 0.072)	(-0.152 - 0.068)	(0.023 - 0.4)	(-0.212 - 0.178)	(-0.047 - 0.016)	(-0.007 - 0.071)	(-0.219 - 0.18)

Table B.30: Performance metrics across disparity groups in the test set with XGBoost, in the Thermometry study.

Task	Feature		AUROC				Recall				F1-score				Accuracy			
			(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)	(-∞, -1)	[-1, 0)	[0, 1)	[1, +∞)
In-hospital mortality	R		0.780	0.769	0.774	0.792	0.487	0.422	0.454	0.581	0.413	0.442	0.464	0.547	0.761	0.760	0.767	0.738
	T		0.712	0.774	0.755	0.813	0.490	0.434	0.406	0.608	0.446	0.459	0.437	0.591	0.764	0.766	0.773	0.763
		P-value	0.165	0.592	0.204	0.019	0.954	0.628	0.038	0.183	0.557	0.423	0.096	0.242	0.910	0.561	0.616	0.310
		95% CI	(-0.034 - 0.171)	(-0.028 - 0.017)	(-0.012 - 0.049)	(-0.038 - - 0.004)	(-0.129 - 0.122)	(-0.067 - 0.043)	(0.003 - 0.092)	(-0.069 - 0.015)	(-0.159 - 0.092)	(-0.064 - 0.029)	(-0.006 - 0.06)	(-0.121 - 0.035)	(-0.053 - 0.048)	(-0.029 - 0.017)	(-0.029 - 0.018)	(-0.078 - 0.028)
	C1		0.734	0.805	0.785	0.735	0.292	0.478	0.458	0.457	0.213	0.471	0.487	0.365	0.603	0.763	0.784	0.694
		P-value	0.675	0.080	0.438	0.769	0.196	0.471	0.898	0.510	0.219	0.631	0.454	0.239	0.071	0.888	0.358	0.639
		95% CI	(-0.195 - 0.287)	(-0.079 - 0.005)	(-0.042 - 0.02)	(-0.203 - 0.263)	(-0.121 - 0.511)	(-0.223 - 0.112)	(-0.074 - 0.066)	(-0.286 - 0.534)	(-0.142 - 0.541)	(-0.158 - 0.101)	(-0.088 - 0.043)	(-0.145 - 0.511)	(-0.017 - 0.333)	(-0.05 - 0.044)	(-0.057 - 0.023)	(-0.161 - 0.249)
	C2		0.768	0.792	0.776	0.788	0.392	0.464	0.428	0.323	0.371	0.472	0.462	0.285	0.702	0.782	0.762	0.663
		P-value	0.906	0.353	0.924	0.969	0.589	0.429	0.531	0.124	0.827	0.550	0.962	0.065	0.409	0.281	0.806	0.444
		95% CI	(-0.222 - 0.248)	(-0.076 - 0.03)	(-0.046 - 0.042)	(-0.261 - 0.27)	(-0.289 - 0.48)	(-0.156 - 0.072)	(-0.064 - 0.116)	(-0.086 - 0.602)	(-0.381 - 0.465)	(-0.137 - 0.078)	(-0.081 - 0.085)	(-0.02 - 0.545)	(-0.094 - 0.212)	(-0.065 - 0.021)	(-0.042 - 0.053)	(-0.137 - 0.287)
	C3		0.810	0.785	0.777	0.722	0.303	0.469	0.437	0.631	0.246	0.457	0.476	0.529	0.629	0.788	0.764	0.652
		P-value	0.588	0.473	0.743	0.489	0.121	0.515	0.710	0.714	0.051	0.825	0.684	0.844	0.065	0.206	0.886	0.350
		95% CI	(-0.278 - 0.169)	(-0.066 - 0.033)	(-0.025 - 0.018)	(-0.14 - 0.269)	(-0.059 - 0.427)	(-0.204 - 0.11)	(-0.083 - 0.117)	(-0.349 - 0.249)	(-0.001 - 0.333)	(-0.167 - 0.137)	(-0.079 - 0.054)	(-0.186 - 0.223)	(-0.01 - 0.275)	(-0.076 - 0.019)	(-0.045 - 0.051)	(-0.111 - 0.283)
Respiratory SOFA score	R		0.839	0.893	0.880	0.850	0.718	0.764	0.707	0.683	0.692	0.780	0.711	0.675	0.832	0.846	0.847	0.785
	T		0.860	0.891	0.886	0.825	0.701	0.761	0.702	0.655	0.699	0.782	0.728	0.656	0.838	0.847	0.862	0.778
		P-value	0.146	0.712	0.435	0.328	0.636	0.919	0.682	0.729	0.801	0.933	0.106	0.801	0.826	0.931	0.027	0.846
		95% CI	(-0.009)	(-0.006 - 0.008)	(-0.024 - 0.011)	(-0.029 - 0.079)	(-0.063 - 0.098)	(-0.064 - 0.07)	(-0.023 - 0.034)	(-0.151 - 0.207)	(-0.074 - 0.059)	(-0.051 - 0.048)	(-0.04 - 0.005)	(-0.142 - 0.179)	(-0.064 - 0.052)	(-0.032 - 0.029)	(-0.028 - 0.002)	(-0.076 - 0.091)
	C1		0.910	0.892	0.886	0.836	0.712	0.749	0.738	0.625	0.706	0.794	0.751	0.628	0.838	0.860	0.847	0.812
		P-value	0.429	0.988	0.666	0.584	0.976	0.721	0.328	0.657	0.942	0.709	0.129	0.721	0.936	0.536	0.977	0.664
		95% CI	(-0.24 - 0.113)	(-0.045 - 0.045)	(-0.036 - 0.024)	(-0.144 - 0.236)	(-0.446 - 0.459)	(-0.079 - 0.11)	(-0.098 - 0.036)	(-0.229 - 0.345)	(-0.438 - 0.41)	(-0.092 - 0.065)	(-0.095 - 0.014)	(-0.241 - 0.334)	(-0.17 - 0.158)	(-0.064 - 0.035)	(-0.023 - 0.022)	(-0.161 - 0.108)
	C2		0.853	0.887	0.887	0.772	0.708	0.744	0.748	0.533	0.716	0.772	0.753	0.532	0.845	0.854	0.847	0.789
		P-value	0.947	0.614	0.639	0.292	0.963	0.399	0.073	0.351	0.899	0.731	0.105	0.385	0.864	0.544	0.981	0.951
		95% CI	(-0.217)	(-0.032)	(-0.027)	(-0.105 - 0.294)	(-0.44 - 0.459)	(-0.031 - 0.072)	(-0.087 - 0.005)	(-0.195 - 0.495)	(-0.448 - 0.399)	(-0.047 - 0.064)	(-0.095 - 0.011)	(-0.21 - 0.494)	(-0.18 - 0.154)	(-0.038 - 0.021)	(-0.029 - 0.028)	(-0.159 - 0.15)
	C3		0.866	0.875	0.900	0.833	0.780	0.717	0.746	0.792	0.825	0.731	0.760	0.686	0.842	0.840	0.852	0.740
		P-value	0.864	0.204	0.171	0.821	0.659	0.030	0.374	0.265	0.270	0.035	0.195	0.887	0.845	0.643	0.790	0.447
		95% CI	(-0.224 - 0.193)	(-0.012 - 0.047)	(-0.05 - 0.01)	(-0.145 - 0.179)	(-0.368 - 0.244)	(0.006 - 0.089)	(-0.132 - 0.055)	(-0.315 - 0.098)	(-0.39 - 0.123)	(0.004 - 0.095)	(-0.129 - 0.03)	(-0.19 - 0.167)	(-0.12 - 0.101)	(-0.02 - 0.031)	(-0.042 - 0.033)	(-0.084 - 0.174)
Overall SOFA score increase	R		0.781	0.655	0.682	0.548	0.312	0.339	0.364	0.335	0.366	0.385	0.400	0.379	0.782	0.728	0.735	0.658
	T		0.767	0.657	0.681	0.579	0.587	0.409	0.337	0.256	0.554	0.419	0.384	0.318	0.801	0.710	0.745	0.680
		P-value	0.664	0.848	0.892	0.442	0.056	0.107	0.232	0.027	0.125	0.370	0.462	0.125	0.454	0.249	0.577	0.348
		95% CI	(-0.081)	(-0.019 - 0.016)	(-0.024 - 0.027)	(-0.122 - 0.058)	(-0.558 - 0.008)	(-0.158 - 0.018)	(-0.021 - 0.074)	(0.011 - 0.146)	(-0.441 - 0.064)	(-0.115 - 0.047)	(-0.03 - 0.062)	(-0.021 - 0.144)	(-0.072 - 0.035)	(-0.015 - 0.052)	(-0.048 - 0.029)	(-0.072 - 0.028)
	C1		0.696	0.662	0.661	0.819	0.325	0.332	0.350	0.207	0.340	0.363	0.393	0.267	0.769	0.728	0.721	0.721
		P-value	0.629	0.909	0.208	0.169	0.950	0.932	0.692	0.206	0.907	0.813	0.841	0.326	0.898	0.990	0.410	0.352
		95% CI	(-0.205 - 0.313)	(-0.134 - 0.121)	(-0.015 - 0.058)	(-0.652 - 0.15)	(-0.476 - 0.449)	(-0.187 - 0.203)	(-0.062 - 0.09)	(-0.084 - 0.34)	(-0.457 - 0.509)	(-0.182 - 0.226)	(-0.069 - 0.083)	(-0.132 - 0.356)	(-0.212 - 0.239)	(-0.087 - 0.088)	(-0.023 - 0.052)	(-0.208 - 0.082)
	C2		0.674	0.690	0.678	0.855	0.475	0.346	0.371	0.180	0.400	0.388	0.418	0.219	0.730	0.719	0.745	0.632
		P-value	0.325	0.316	0.888	0.115	0.271	0.903	0.882	0.074	0.833	0.950	0.698	0.124	0.455	0.691	0.726	0.717
		95% CI	(-0.132 - 0.345)	(-0.109 - 0.039)	(-0.064 - 0.073)	(-0.826 - 0.132)	(-0.478 - 0.152)	(-0.13 - 0.116)	(-0.103 - 0.09)	(-0.019 - 0.328)	(-0.389 - 0.321)	(-0.125 - 0.118)	(-0.122 - 0.085)	(-0.053 - 0.374)	(-0.099 - 0.204)	(-0.042 - 0.061)	(-0.071 - 0.052)	(-0.133 - 0.185)
	C3		0.649	0.647	0.688	0.741	0.250	0.328	0.325	0.098	0.247	0.352	0.391	0.154	0.785	0.725	0.726	0.657
		P-value	0.198	0.871	0.828	0.170	0.775	0.866	0.155	0.040	0.569	0.628	0.801	0.097	0.973	0.901	0.584	0.986
		95% CI	(-0.076 - 0.296)	(-0.098 - 0.113)	(-0.061 - 0.05)	(-0.411 - 0.091)	(-0.412 - 0.536)	(-0.135 - 0.157)	(-0.018 - 0.096)	(0.014 - 0.46)	(-0.336 - 0.574)	(-0.115 - 0.18)	(-0.063 - 0.08)	(-0.05 - 0.501)	(-0.201 - 0.195)	(-0.049 - 0.054)	(-0.03 - 0.05)	(-0.127 - 0.129)

Table B.31: Performance metrics in the test set by hidden hypothermia, normothermia and hidden fever, with LR.

Task	Feature		AUROC			Recall			F1-score			Accuracy		
			Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever
In-hospital mortality	R		0.894	0.790	0.788	0.782	0.706	0.508	0.724	0.512	0.485	0.776	0.722	0.755
	T		0.894	0.793	0.777	0.782	0.722	0.508	0.729	0.524	0.462	0.784	0.729	0.739
		P-value		0.188	0.431		0.207		0.343	0.116	0.343	0.343	0.248	0.343
		95% CI	(0.0 - 0.0)	(-0.009 - 0.002)	(-0.018 - 0.039)	(0.0 - 0.0)	(-0.042 - 0.01)	(0.0 - 0.0)	(-0.014 - 0.006)	(-0.028 - 0.004)	(-0.029 - 0.075)	(-0.025 - 0.01)	(-0.018 - 0.005)	(-0.019 - 0.05)
	C1	P-value	1.000	0.796	0.881	0.500	0.727	0.450	0.500	0.514	0.433	0.875	0.725	0.681
		95% CI	(nan - nan)	0.487	0.172	0.241	0.384	0.702	0.334	0.828	0.701	0.414	0.536	0.392
				(-0.026 - 0.013)	(-0.316 - 0.071)	(-0.227 - 0.791)	(-0.073 - 0.031)	(-0.275 - 0.392)	(-0.273 - 0.721)	(-0.023 - 0.019)	(-0.242 - 0.345)	(-0.499 - 0.231)	(-0.013 - 0.007)	(-0.112 - 0.26)
	C2	P-value	1.000	0.795	0.888	0.450	0.726	0.450	0.467	0.513	0.433	0.952	0.724	0.681
		95% CI	(nan - nan)	0.522	0.165	0.165	0.350	0.702	0.258	0.927	0.701	0.037	0.822	0.392
				(-0.024 - 0.013)	(-0.329 - 0.071)	(-0.165 - 0.83)	(-0.066 - 0.026)	(-0.275 - 0.392)	(-0.225 - 0.74)	(-0.017 - 0.016)	(-0.242 - 0.345)	(-0.422 - -0.018)	(-0.012 - 0.01)	(-0.112 - 0.26)
	C3	P-value	0.845	0.798	0.733	0.817	0.725	0.506	0.741	0.515	0.398	0.845	0.725	0.646
		95% CI	0.815	0.282	0.428	0.837	0.367	0.984	0.917	0.719	0.416	0.437	0.616	<.01
			(-0.163 - 0.2)	(-0.025 - 0.008)	(-0.108 - 0.23)	(-0.399 - 0.33)	(-0.065 - 0.026)	(-0.29 - 0.295)	(-0.372 - 0.339)	(-0.023 - 0.016)	(-0.143 - 0.316)	(-0.26 - 0.123)	(-0.013 - 0.008)	(0.048 - 0.17)
Respiratory SOFA score	R		0.867	0.870	0.806	0.810	0.756	0.648	0.776	0.713	0.605	0.906	0.835	0.801
	T		0.882	0.866	0.803	0.798	0.736	0.743	0.753	0.702	0.637	0.876	0.828	0.779
		P-value	0.449	0.248	0.711	0.343	0.167	0.112	0.209	0.224	0.504	0.193	0.234	0.632
		95% CI	(-0.059 - 0.029)	(-0.003 - 0.011)	(-0.017 - 0.024)	(-0.016 - 0.041)	(-0.01 - 0.048)	(-0.217 - 0.027)	(-0.016 - 0.063)	(-0.008 - 0.031)	(-0.137 - 0.072)	(-0.018 - 0.078)	(-0.005 - 0.019)	(-0.077 - 0.121)
	C1	P-value		0.867	0.888	0.400	0.731	0.667	0.400	0.718	0.597	0.938	0.829	0.792
		95% CI	(nan - nan)	0.621	0.228	0.049	0.288	0.906	0.043	0.732	0.943	0.675	0.284	0.886
				(-0.01 - 0.016)	(-0.247 - 0.069)	(0.002 - 0.818)	(-0.024 - 0.073)	(-0.358 - 0.322)	(0.014 - 0.738)	(-0.038 - 0.028)	(-0.24 - 0.256)	(-0.154 - 0.106)	(-0.007 - 0.02)	(-0.124 - 0.141)
	C2	P-value		0.864	0.847	0.400	0.723	0.633	0.400	0.715	0.547	0.929	0.826	0.768
		95% CI	(nan - nan)	0.283	0.284	0.049	0.121	0.931	0.043	0.941	0.660	0.680	0.151	0.681
				(-0.006 - 0.019)	(-0.147 - 0.049)	(0.002 - 0.818)	(-0.011 - 0.076)	(-0.368 - 0.399)	(0.014 - 0.738)	(-0.034 - 0.032)	(-0.231 - 0.347)	(-0.182 - 0.127)	(-0.004 - 0.021)	(-0.145 - 0.213)
	C3	P-value	0.955	0.861	0.861	0.800	0.718	0.719	0.663	0.702	0.664	0.845	0.823	0.774
		95% CI	0.181	0.224	0.440	0.960	0.131	0.695	0.502	0.517	0.666	0.367	0.091	0.667
			(-0.195 - 0.046)	(-0.007 - 0.025)	(-0.21 - 0.099)	(-0.43 - 0.45)	(-0.014 - 0.09)	(-0.467 - 0.325)	(-0.252 - 0.478)	(-0.026 - 0.048)	(-0.356 - 0.239)	(-0.084 - 0.206)	(-0.002 - 0.027)	(-0.11 - 0.164)
Overall SOFA score increase	R		0.732	0.680	0.586	0.455	0.584	0.250	0.422	0.435	0.140	0.684	0.634	0.671
	T		0.742	0.680	0.497	0.505	0.590	0.350	0.462	0.433	0.180	0.701	0.627	0.698
		P-value	0.787	0.871	0.351	0.343	0.556	0.343	0.343	0.762	0.343	0.343	0.172	0.096
		95% CI	(-0.097 - 0.077)	(-0.004 - 0.004)	(-0.118 - 0.296)	(-0.163 - 0.063)	(-0.028 - 0.016)	(-0.326 - 0.126)	(-0.13 - 0.05)	(-0.013 - 0.017)	(-0.13 - 0.05)	(-0.054 - 0.021)	(-0.004 - 0.018)	(-0.06 - 0.006)
	C1	P-value	0.000	0.693	0.293	0.100	0.617	0.200	0.100	0.452	0.133	0.833	0.634	0.657
		95% CI	0.144	0.108	0.540	0.036	<.01	0.823	0.037	0.069	0.962	0.220	0.989	0.862
			(-1.256 - 2.556)	(-0.029 - 0.003)	(-0.507 - 0.829)	(0.03 - 0.68)	(-0.055 - -0.012)	(-0.44 - 0.54)	(0.025 - 0.619)	(-0.035 - 0.002)	(-0.302 - 0.315)	(-0.54 - 0.159)	(-0.016 - 0.016)	(-0.164 - 0.192)
	C2	P-value	0.000	0.700	0.533	0.000	0.624	0.200	0.000	0.454	0.133	0.800	0.633	0.753
		95% CI	0.144	0.030	0.726	0.010	<.01	0.823	0.010	0.073	0.962	0.300	0.781	0.286
			(-1.256 - 2.556)	(-0.038 - -0.002)	(-0.662 - 0.504)	(0.138 - 0.771)	(-0.064 - -0.016)	(-0.44 - 0.54)	(0.128 - 0.716)	(-0.04 - 0.002)	(-0.302 - 0.315)	(-0.66 - 0.264)	(-0.012 - 0.016)	(-0.247 - 0.082)
	C3	P-value	0.731	0.701	0.554	0.452	0.609	0.242	0.522	0.454	0.224	0.723	0.638	0.695
		95% CI	0.996	<.01	0.736	0.988	0.077	0.963	0.602	0.175	0.484	0.590	0.739	0.705
			(-0.33 - 0.331)	(-0.036 - -0.008)	(-0.247 - 0.333)	(-0.426 - 0.432)	(-0.055 - 0.003)	(-0.382 - 0.398)	(-0.518 - 0.318)	(-0.047 - 0.01)	(-0.343 - 0.176)	(-0.198 - 0.119)	(-0.027 - 0.02)	(-0.168 - 0.119)

Table B.32: Performance metrics in the test set by hidden hypothermia, normothermia and hidden fever, with RF.

Task	Feature		AUROC			Recall			F1-score			Accuracy		
			Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever
In-hospital mortality	R		0.889	0.789	0.827	0.320	0.242	0.160	0.374	0.333	0.175	0.687	0.806	0.750
	T		0.722	0.778	0.813	0.306	0.272	0.160	0.352	0.349	0.192	0.679	0.801	0.739
		P-value	0.077	0.132	0.689	0.343	0.151	1.000	0.343	0.486	0.906	0.343	0.545	0.554
		95% CI	(-0.023 - 0.355)	(-0.004 - 0.026)	(-0.065 - 0.094)	(-0.018 - 0.047)	(-0.074 - 0.013)	(-0.292 - 0.292)	(-0.028 - 0.072)	(-0.064 - 0.033)	(-0.327 - 0.293)	(-0.01 - 0.025)	(-0.014 - 0.025)	(-0.031 - 0.054)
	C1		1.000	0.794	0.896	0.280	0.269	0.000	0.289	0.350	0.000	0.767	0.808	0.655
		P-value		0.682	0.211	0.831	0.086	0.182	0.682	0.424	0.173	0.688	0.840	0.351
		95% CI	(nan - nan)	(-0.029 - 0.02)	(-0.258 - 0.07)	(-0.372 - 0.452)	(-0.057 - 0.005)	(-0.091 - 0.411)	(-0.371 - 0.542)	(-0.064 - 0.029)	(-0.092 - 0.442)	(-0.495 - 0.346)	(-0.023 - 0.019)	(-0.124 - 0.314)
	C2		1.000	0.796	0.881	0.180	0.237	0.000	0.189	0.314	0.000	0.590	0.802	0.655
		P-value		0.504	0.396	0.349	0.682	0.182	0.285	0.306	0.173	0.716	0.677	0.351
		95% CI	(nan - nan)	(-0.031 - 0.017)	(-0.29 - 0.132)	(-0.18 - 0.461)	(-0.025 - 0.037)	(-0.091 - 0.411)	(-0.184 - 0.555)	(-0.021 - 0.06)	(-0.092 - 0.442)	(-0.467 - 0.64)	(-0.017 - 0.025)	(-0.124 - 0.314)
	C3		1.000	0.786	0.701	0.350	0.239	0.161	0.367	0.310	0.228	0.629	0.796	0.689
		P-value	0.025	0.701	0.193	0.868	0.904	0.991	0.967	0.514	0.580	0.686	0.379	0.375
		95% CI	(-0.196 - -0.019)	(-0.012 - 0.017)	(-0.083 - 0.348)	(-0.426 - 0.366)	(-0.055 - 0.061)	(-0.19 - 0.188)	(-0.405 - 0.42)	(-0.054 - 0.1)	(-0.263 - 0.156)	(-0.256 - 0.372)	(-0.014 - 0.034)	(-0.087 - 0.21)
Respiratory SOFA score	R		0.939	0.898	0.844	0.750	0.680	0.635	0.774	0.732	0.666	0.901	0.865	0.860
	T		0.939	0.892	0.876	0.709	0.672	0.577	0.749	0.734	0.570	0.852	0.868	0.840
		P-value		0.138	0.048	0.430	0.247	0.399	0.480	0.768	0.273	0.206	0.388	0.356
		95% CI	(0.0 - 0.0)	(-0.002 - 0.014)	(-0.065 - -0.0)	(-0.071 - 0.152)	(-0.006 - 0.021)	(-0.09 - 0.207)	(-0.051 - 0.1)	(-0.017 - 0.013)	(-0.09 - 0.283)	(-0.032 - 0.13)	(-0.012 - 0.005)	(-0.027 - 0.067)
	C1		0.896	0.870	0.400	0.677	0.417	0.400	0.734	0.413	1.000	0.855	0.725	
		P-value		0.735	0.433	0.103	0.839	0.096	0.065	0.850	0.014	0.093	0.220	0.157
		95% CI	(nan - nan)	(-0.009 - 0.012)	(-0.078 - 0.037)	(-0.086 - 0.786)	(-0.026 - 0.031)	(-0.048 - 0.484)	(-0.028 - 0.776)	(-0.032 - 0.027)	(0.064 - 0.441)	(-0.229 - 0.022)	(-0.007 - 0.026)	(-0.063 - 0.333)
	C2		0.899	0.944	0.400	0.692	0.617	0.400	0.747	0.627	1.000	0.861	0.833	
		P-value		0.873	0.282	0.103	0.351	0.906	0.065	0.132	0.702	0.092	0.554	0.773
		95% CI	(nan - nan)	(-0.012 - 0.01)	(-0.284 - 0.095)	(-0.086 - 0.786)	(-0.042 - 0.016)	(-0.323 - 0.36)	(-0.028 - 0.776)	(-0.037 - 0.006)	(-0.187 - 0.265)	(-0.262 - 0.026)	(-0.009 - 0.015)	(-0.178 - 0.231)
	C3		0.955	0.902	0.784	0.600	0.702	0.618	0.580	0.758	0.662	0.885	0.870	0.829
		P-value	0.239	0.456	0.600	0.469	0.211	0.904	0.286	0.125	0.965	0.793	0.468	0.490
		95% CI	(-0.089 - 0.027)	(-0.016 - 0.008)	(-0.19 - 0.31)	(-0.299 - 0.599)	(-0.059 - 0.015)	(-0.292 - 0.326)	(-0.193 - 0.581)	(-0.061 - 0.009)	(-0.227 - 0.236)	(-0.116 - 0.148)	(-0.023 - 0.011)	(-0.067 - 0.129)
Overall SOFA score increase	R		0.826	0.678	0.575	0.100	0.086	0.000	0.133	0.145	0.000	0.708	0.767	0.848
	T		0.687	0.699	0.633	0.100	0.080	0.100	0.133	0.137	0.100	0.688	0.768	0.839
		P-value	0.040	0.208	0.656		0.609	0.343		0.689	0.343	0.343	0.863	0.523
		95% CI	(0.009 - 0.268)	(-0.054 - 0.014)	(-0.345 - 0.23)	(0.0 - 0.0)	(-0.019 - 0.03)	(-0.326 - 0.126)	(0.0 - 0.0)	(-0.035 - 0.05)	(-0.326 - 0.126)	(-0.025 - 0.065)	(-0.01 - 0.008)	(-0.021 - 0.039)
	C1		0.500	0.680	0.925	0.000	0.097	0.000	0.000	0.165	0.000	0.708	0.768	0.872
		P-value	0.656	0.924	0.039	0.173	0.422		0.170	0.382		0.844	0.784	0.690
		95% CI	(-7.566 - 8.316)	(-0.029 - 0.027)	(-0.91 - -0.041)	(-0.053 - 0.253)	(-0.041 - 0.019)	(0.0 - 0.0)	(-0.069 - 0.334)	(-0.069 - 0.029)	(0.0 - 0.0)	(-0.391 - 0.459)	(-0.012 - 0.009)	(-0.153 - 0.106)
	C2		1.000	0.669	0.850	0.000	0.105	0.000	0.000	0.178	0.000	0.850	0.769	0.868
		P-value	0.500	0.405	0.099	0.173	0.251		0.170	0.232		0.472	0.626	0.736
		95% CI	(-1.713 - 1.463)	(-0.015 - 0.033)	(-0.92 - 0.119)	(-0.053 - 0.253)	(-0.055 - 0.016)	(0.0 - 0.0)	(-0.069 - 0.334)	(-0.09 - 0.025)	(0.0 - 0.0)	(-0.438 - 0.243)	(-0.014 - 0.009)	(-0.151 - 0.111)
	C3		0.676	0.691	0.670	0.000	0.104	0.133	0.000	0.176	0.150	0.588	0.766	0.824
		P-value	0.246	0.150	0.285	0.173	0.241	0.223	0.170	0.228	0.193	0.290	0.896	0.649
		95% CI	(-0.151 - 0.466)	(-0.031 - 0.006)	(-0.452 - 0.155)	(-0.053 - 0.253)	(-0.052 - 0.015)	(-0.364 - 0.097)	(-0.069 - 0.334)	(-0.084 - 0.023)	(-0.391 - 0.091)	(-0.121 - 0.362)	(-0.011 - 0.012)	(-0.092 - 0.14)

Table B.33: Performance metrics in the test set by hidden hypothermia, normothermia and hidden fever, with XGBoost.

Task	Feature		AUROC			Recall			F1-score			Accuracy		
			Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever	Hidden Hypothermia	Normothermia	Hidden Fever
In-hospital mortality	R		0.865	0.770	0.775	0.675	0.414	0.435	0.664	0.431	0.479	0.816	0.776	0.804
	T		0.750	0.763	0.767	0.548	0.412	0.410	0.574	0.435	0.459	0.768	0.781	0.791
		P-value	0.342	0.492	0.772	0.238	0.950	0.537	0.207	0.791	0.641	0.091	0.324	0.455
		95% CI	(-0.146 - 0.374)	(-0.016 - 0.031)	(-0.054 - 0.07)	(-0.1 - 0.355)	(-0.044 - 0.047)	(-0.063 - 0.113)	(-0.059 - 0.239)	(-0.04 - 0.031)	(-0.075 - 0.115)	(-0.009 - 0.105)	(-0.016 - 0.006)	(-0.025 - 0.052)
	C1		1.000	0.777	0.922	0.480	0.416	0.250	0.489	0.431	0.267	0.975	0.782	0.737
		P-value		0.599	0.049	0.478	0.946	0.296	0.506	0.988	0.263	0.053	0.586	0.284
		95% CI	(nan - nan)	(-0.033 - 0.02)	(-0.28 - -0.001)	(-0.401 - 0.792)	(-0.057 - 0.053)	(-0.192 - 0.562)	(-0.395 - 0.745)	(-0.059 - 0.06)	(-0.19 - 0.616)	(-0.372 - 0.003)	(-0.032 - 0.019)	(-0.066 - 0.199)
	C2		1.000	0.777	0.935	0.380	0.402	0.150	0.389	0.418	0.167	0.829	0.778	0.712
		P-value		0.536	0.045	0.309	0.599	0.026	0.325	0.553	0.030	0.846	0.868	0.206
		95% CI	(nan - nan)	(-0.029 - 0.016)	(-0.303 - -0.005)	(-0.325 - 0.915)	(-0.037 - 0.061)	(0.043 - 0.527)	(-0.322 - 0.872)	(-0.036 - 0.063)	(0.038 - 0.588)	(-0.513 - 0.434)	(-0.024 - 0.021)	(-0.061 - 0.246)
	C3		0.821	0.770	0.750	0.650	0.425	0.278	0.630	0.436	0.355	0.798	0.780	0.745
		P-value	0.965	0.976	0.512	0.909	0.567	0.129	0.859	0.826	0.285	0.817	0.749	0.126
		95% CI	(-0.234 - 0.243)	(-0.022 - 0.022)	(-0.107 - 0.198)	(-0.462 - 0.513)	(-0.052 - 0.03)	(-0.056 - 0.37)	(-0.38 - 0.447)	(-0.058 - 0.048)	(-0.123 - 0.372)	(-0.156 - 0.193)	(-0.031 - 0.023)	(-0.02 - 0.137)
Respiratory SOFA score	R		0.959	0.882	0.820	0.776	0.738	0.718	0.742	0.735	0.666	0.862	0.855	0.817
	T		0.936	0.883	0.790	0.726	0.719	0.707	0.766	0.740	0.697	0.872	0.865	0.864
		P-value	0.141	0.724	0.497	0.343	0.399	0.343	0.719	0.776	0.303	0.883	0.332	0.288
		95% CI	(-0.01 - 0.057)	(-0.012 - 0.008)	(-0.065 - 0.123)	(-0.063 - 0.163)	(-0.03 - 0.069)	(-0.014 - 0.036)	(-0.171 - 0.123)	(-0.046 - 0.036)	(-0.093 - 0.032)	(-0.151 - 0.132)	(-0.031 - 0.012)	(-0.142 - 0.048)
	C1			0.889	0.752	0.400	0.737	0.617	0.400	0.753	0.586	1.000	0.856	0.800
		P-value		0.089	0.429	0.095	0.931	0.524	0.080	0.127	0.554	0.047	0.911	0.868
		95% CI	(nan - nan)	(-0.017 - 0.001)	(-0.133 - 0.283)	(-0.08 - 0.832)	(-0.032 - 0.034)	(-0.245 - 0.449)	(-0.05 - 0.733)	(-0.042 - 0.006)	(-0.216 - 0.377)	(-0.28 - -0.002)	(-0.015 - 0.013)	(-0.207 - 0.24)
	C2			0.886	0.833	0.400	0.747	0.617	0.400	0.761	0.627	1.000	0.859	0.833
		P-value		0.336	0.934	0.095	0.409	0.524	0.080	0.031	0.777	0.045	0.578	0.875
		95% CI	(nan - nan)	(-0.014 - 0.005)	(-0.161 - 0.15)	(-0.08 - 0.832)	(-0.033 - 0.015)	(-0.245 - 0.449)	(-0.05 - 0.733)	(-0.048 - -0.003)	(-0.268 - 0.348)	(-0.317 - -0.005)	(-0.018 - 0.011)	(-0.245 - 0.212)
	C3		0.913	0.887	0.768	0.550	0.722	0.599	0.513	0.739	0.578	0.860	0.852	0.771
		P-value	0.605	0.248	0.674	0.253	0.507	0.490	0.125	0.814	0.568	0.971	0.575	0.363
		95% CI	(-0.108 - 0.169)	(-0.016 - 0.005)	(-0.219 - 0.323)	(-0.192 - 0.643)	(-0.036 - 0.068)	(-0.255 - 0.493)	(-0.077 - 0.533)	(-0.038 - 0.031)	(-0.248 - 0.424)	(-0.124 - 0.128)	(-0.01 - 0.017)	(-0.063 - 0.155)
Overall SOFA score increase	R		0.646	0.667	0.659	0.219	0.343	0.100	0.265	0.380	0.067	0.617	0.733	0.820
	T		0.720	0.665	0.654	0.205	0.380	0.150	0.253	0.402	0.117	0.712	0.730	0.783
		P-value	0.356	0.920	0.924	0.343	0.274	0.343	0.343	0.500	0.343	0.190	0.869	0.454
		95% CI	(-0.253 - 0.106)	(-0.029 - 0.032)	(-0.104 - 0.114)	(-0.018 - 0.046)	(-0.108 - 0.035)	(-0.163 - 0.063)	(-0.016 - 0.041)	(-0.09 - 0.048)	(-0.163 - 0.063)	(-0.247 - 0.057)	(-0.034 - 0.04)	(-0.069 - 0.143)
	C1		0.000	0.653	0.520	0.000	0.338	0.200	0.000	0.366	0.200	0.708	0.731	0.880
		P-value	0.249	0.391	0.665	0.053	0.906	0.591	0.044	0.662	0.423	0.757	0.881	0.438
		95% CI	(-2.4 - 3.533)	(-0.021 - 0.048)	(-0.583 - 0.819)	(-0.004 - 0.442)	(-0.084 - 0.094)	(-0.506 - 0.306)	(0.009 - 0.521)	(-0.057 - 0.086)	(-0.492 - 0.226)	(-0.553 - 0.428)	(-0.027 - 0.031)	(-0.228 - 0.108)
	C2		0.334	0.674	0.800	0.100	0.345	0.300	0.067	0.386	0.267	0.550	0.743	0.910
		P-value	0.752	0.685	0.268	0.467	0.971	0.343	0.205	0.887	0.260	0.777	0.617	0.258
		95% CI	(-6.971 - 7.437)	(-0.045 - 0.031)	(-0.513 - 0.188)	(-0.235 - 0.474)	(-0.085 - 0.083)	(-0.652 - 0.252)	(-0.131 - 0.528)	(-0.085 - 0.075)	(-0.576 - 0.176)	(-0.584 - 0.727)	(-0.051 - 0.032)	(-0.259 - 0.079)
	C3		0.675	0.650	0.616	0.200	0.309	0.175	0.207	0.351	0.135	0.670	0.727	0.799
		P-value	0.609	0.196	0.890	0.892	0.157	0.193	0.701	0.183	0.174	0.694	0.560	0.722
		95% CI	(-0.143 - 0.221)	(-0.01 - 0.043)	(-0.283 - 0.32)	(-0.293 - 0.331)	(-0.016 - 0.083)	(-0.196 - 0.046)	(-0.276 - 0.394)	(-0.017 - 0.075)	(-0.174 - 0.037)	(-0.352 - 0.245)	(-0.018 - 0.031)	(-0.105 - 0.146)

Table B.34: Fairness metrics in the test set, in the Thermometry study.

Task	Feature		**LR**		**RF**		**XGBoost**	
			Demographic parity difference	Equalized odds ratio	Demographic parity difference	Equalized odds ratio	Demographic parity difference	Equalized odds ratio
In-hospital mortality	R		0.559	0.009	0.384	0.000	0.472	0.009
	T		0.527	0.009	0.400	0.000	0.444	0.011
		P-value	0.116		0.612		0.444	0.343
		95% CI	(-0.01 - 0.074)	(0.0 - 0.0)	(-0.087 - 0.054)	(0.0 - 0.0)	(-0.051 - 0.106)	(-0.006 - 0.002)
	C1		0.532	0.009	0.420	0.000	0.481	0.000
		P-value	0.034		0.179		0.768	0.343
		95% CI	(0.003 - 0.051)	(0.0 - 0.0)	(-0.092 - 0.02)	(0.0 - 0.0)	(-0.076 - 0.058)	(-0.011 - 0.029)
	C2		0.532	0.009	0.389	0.000	0.500	0.004
		P-value	0.080		0.873		0.485	0.343
		95% CI	(-0.004 - 0.057)	(0.0 - 0.0)	(-0.084 - 0.073)	(0.0 - 0.0)	(-0.114 - 0.059)	(-0.006 - 0.015)
	C3		0.551	0.009	0.399	0.000	0.458	0.007
		P-value	0.178		0.647		0.789	0.343
		95% CI	(-0.005 - 0.022)	(0.0 - 0.0)	(-0.091 - 0.06)	(0.0 - 0.0)	(-0.096 - 0.122)	(-0.003 - 0.007)
Respiratory SOFA score	R		0.540	0.062	0.542	0.023	0.510	0.025
	T		0.465	0.057	0.490	0.015	0.515	0.016
		P-value	0.088	0.206	0.384	0.343	0.913	0.343
		95% CI	(-0.014 - 0.163)	(-0.003 - 0.014)	(-0.077 - 0.181)	(-0.01 - 0.025)	(-0.096 - 0.087)	(-0.011 - 0.029)
	C1		0.515	0.025	0.481	0.026	0.527	0.015
		P-value	0.232	0.196	0.356	0.832	0.504	0.343
		95% CI	(-0.019 - 0.07)	(-0.023 - 0.097)	(-0.08 - 0.202)	(-0.034 - 0.028)	(-0.071 - 0.038)	(-0.013 - 0.033)
	C2		0.519	0.050	0.492	0.023	0.534	0.020
		P-value	0.528	0.343	0.415		0.223	0.343
		95% CI	(-0.051 - 0.092)	(-0.015 - 0.039)	(-0.081 - 0.18)	(0.0 - 0.0)	(-0.065 - 0.017)	(-0.006 - 0.017)
	C3		0.523	0.025	0.531	0.010	0.558	0.013
		P-value	0.621	0.196	0.245	0.343	0.168	0.343
		95% CI	(-0.057 - 0.091)	(-0.023 - 0.097)	(-0.009 - 0.032)	(-0.016 - 0.041)	(-0.119 - 0.024)	(-0.016 - 0.041)
Overall SOFA score increase	R		0.526	0.068	0.152	0.000	0.246	0.000
	T		0.498	0.091	0.126	0.000	0.299	0.000
		P-value	0.086	0.181	0.350		0.262	
		95% CI	(-0.005 - 0.061)	(-0.059 - 0.013)	(-0.034 - 0.086)	(0.0 - 0.0)	(-0.154 - 0.047)	(0.0 - 0.0)
	C1		0.518	0.068	0.130	0.000	0.377	0.000
		P-value	0.181		0.441		0.038	
		95% CI	(-0.004 - 0.019)	(0.0 - 0.0)	(-0.04 - 0.084)	(0.0 - 0.0)	(-0.253 - -0.009)	(0.0 - 0.0)
	C2		0.502	0.088	0.122	0.000	0.312	0.000
		P-value	0.097	0.171	0.259		0.133	
		95% CI	(-0.005 - 0.052)	(-0.05 - 0.01)	(-0.027 - 0.089)	(0.0 - 0.0)	(-0.157 - 0.024)	(0.0 - 0.0)
	C3		0.470	0.099	0.135	0.000	0.348	0.000
		P-value	0.080	0.188	0.509		0.067	
		95% CI	(-0.008 - 0.12)	(-0.08 - 0.018)	(-0.041 - 0.076)	(0.0 - 0.0)	(-0.213 - 0.009)	(0.0 - 0.0)