scientific reports

Check for updates

OPEN

Predictive risk models for COVID-19 patients using the multi-thresholding meta-algorithm

Rosario Delgado[1]✉, Francisco Fernández-Peláez[2], Natàlia Pallarés[3,15], Vicens Diaz-Brito[5], Elisenda Izquierdo[6], Isabel Oriol[4,7,8], Antonella Simonetti[9,13,14], Cristian Tebé[3,4], Sebastià Videla[10,11] & Jordi Carratalà[4,7,12,13]

This study aims to develop a Machine Learning model to assess the risks faced by COVID-19 patients in a hospital setting, focusing specifically on predicting the complications leading to Intensive Care Unit (ICU) admission or mortality, which are minority classes compared to the majority class of discharged patients. We operate within a multiclass framework comprising three distinct classes, and address the challenge of dataset imbalance, a common source of model bias. To effectively manage this, we introduce the Multi-Thresholding meta-algorithm (MTh), an innovative output-level methodology that extends traditional thresholding from binary to multiclass classification. This methodology dynamically adjusts class probabilities using misclassification costs, making it highly effective in imbalanced datasets. Our approach is further enhanced by integrating the simplicity, transparency, and effectiveness of Bayesian networks to create a robust predictive model. Using patient admission data, the model accurately identifies key risk and protective factors for COVID-19 outcomes. Our findings indicate that certain patient characteristics, such as high Charlson Index and pre-existing conditions, significantly influence the risk of ICU admission and mortality. Moreover, we introduce an explanatory model that elucidates the interrelationships among these factors, demonstrating the influence of therapeutic limits on the overall risk assessment of COVID-19 patients. Overall, our research provides a significant contribution to the field of Machine Learning by offering a novel solution for multiclass classification in the context of imbalanced datasets. This model not only enhances predictive accuracy but also supports critical decision-making processes in healthcare, potentially improving patient outcomes and optimizing clinical resource allocation.

Keywords COVID-19 patient risks assessment, Cost-sensitive Machine Learning modelling, Bayesian Networks, Multiclass classification thresholding, Healthcare decision-making

Abbreviations

COVID-19	Coronavirus Disease 2019
ML	Machine Learning
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2: responsible for COVID-19
NBR	Non-rebreather mask: a therapeutic limit
NIMV	Non-invasive mechanical ventilation: a therapeutic limit

[1]Department of Mathematics, Universitat Autònoma de Barcelona, Barcelona, Spain. [2]Applied Artificial Intelligence Unit, Eurecat, Barcelona, Spain. [3]Biostatistics Unit of the Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain. [4]Department of Clinical Sciences, School of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain. [5]Department of Infectious Diseases, Parc Sanitari S. Joan de Deu, Sant Boi de Llobregat, Barcelona, Spain. [6]Department of Anaesthesiology, Viladecans Hospital, Barcelona, Spain. [7]Bellvitge Biomedical Research Institute, Barcelona, Spain. [8]Unitat Malalties Infeccioses, Servei Medicina Interna, Consorci Sanitari Integral, Barcelona, Spain. [9]Àrea de Recerca, Consorci Sanitari Alt Penedès Garraf, Barcelona, Spain. [10]Department of Clinical Pharmacology, Bellvitge University Hospital, Barcelona, Spain. [11]Department of Pathology and Exp. Therapeutics, School Medicine and Health Sci., University of Barcelona, Barcelona, Spain. [12]Department of Infetious Diseases, Bellvitge University Hospital, Barcelona, Spain. [13]CIBERINFEC, Instituto de Salud Carlos III, Sevilla, Spain. [14]Infectious Disease Unit, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain. [15]Department of Basic Clinical Practice, School of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain. ✉email: Rosario.Delgado@uab.cat

ICU	Intensive Care Unit
icu class	Patients requiring ICU admission
exitus class	Patients passing away without ICU care
discharge class	Patients discharged without ICU care
thresholding	A strategy for improving classification algorithms by adjusting thresholds
MTh	Multi-Thresholding meta-algorithm (introduced in this work)
qRT-PCR	Quantitative Reverse Transcription Polymerase Chain Reaction: laboratory technique to quantity RNA molecules
SMOTE	Synthetic Minority Over-sampling TEchnique: a popular oversampling method in ML
Tomek link	A tool for data preprocessing in ML for classification to reduce noise and class overlap
BOSME	Bayesian network-based Over-Sampling MEthod: an oversampling method in ML for non-continuous data
BN	Bayesian Network: a white-box ML methodology used for classification
Markov Blanket	In a BN, the Markov Blanket of a node is the set of its parents, children, and any other parents of its children
DAG	Directed Acyclic Graph: a finite graph of vertices connected by directed edges, without directed cycles. Used to represent the structure of a BN
MAP	Maximum A Posteriori criterion, used to predict a class as the one maximizing the posterior distribution
Naive Bayes (NB)	A simple BN classifier that assumes independence between features conditioned on the class variable
Augmented Naive Bayes (AN)	An extension of the Naive Bayes classifier that relaxes the independence assumption
Tree Augmented Naive (TAN)	A particular case of AN with a tree structure: each feature depends on its parent feature and on the class
MLE	Maximum Likelihood Estimation: a statistical method for estimating the parameters of a model
NN	Neural Network: a type of ML model inspired by the human brain, consisting of layers of interconnected nodes
SVM	Support Vector Machine: a supervised ML algorithm for classification which finds the optimal hyperplane to separate different classes in the feature space
RF	Random Forest: an ensemble learning method obtained from many decision trees by the Majority Vote criterion
R	A very popular programming language and free software environment for statistical computing
bnlearn	An open-source R package for learning and working with BNs
BIC	Bayesian Information Criterion: used for model selection. Balances goodness-of-fit of the model and complexity
gRain	An open-source R package for handling and making predictions with BNs
mlearning	An R package providing tools for building and evaluating ML models
K-fold cross-validation	Model validation technique that divides the dataset into *K* equally sized folds. The model is trained on $K-1$ folds and tested on the remaining, repeating the process *K* times, each time with a different test set
Confusion Matrix	Table used to evaluate the performance of a classification model by comparing the predicted and observed classes
N	The total number of instances in the validation set
C_{ij}	Number of observed instances in the validation set belonging to class c_j but assigned by the classifier to class c_i
Accuracy	A performance metric for classification models: the proportion of correctly classified instances
Error rate	A performance metric for classification models: the proportion of incorrectly classified instances
BA	Balanced Accuracy: performance metric that accounts for class imbalance by averaging the accuracy of each class
TC	Total Cost: performance metric for classification models that quantifies the overall cost of misclassification based on a predefined cost matrix, taking into account the specific costs of different types of errors
Cost matrix	A matrix used in classification to quantify the costs associated with different types of classification errors
α	Parameter: cost of misclassifying a patient whose class is exitus as discharge. Half if misclassified as icu
β	Parameter: cost of misclassifying a patient whose class is icu as discharge. Half if misclassified as exitus
Shapiro-Wilk	Statistical test used to check normality on the data before applying parametric statistical tests that require it
Student's t-test	Statistical test to compare the means of two groups assuming normality. Paired if the two groups are related

Wilcoxon signed-rank test	Non-parametric statistical test to compare the medians of paired observations. Does not assume normality
p-value	A measure used in statistical hypothesis testing to determine the strength of the evidence against the null hypothesis: the probability of observing the test statistic or a more extreme value, if the null hypothesis is true
adjusted p-values	A method for controlling the family-wise error rate (FWER) when performing multiple hypothesis tests
FWER	Family-wise error rate: the probability of making one or more Type I error among all the hypotheses tested
Type I error	The incorrect rejection of the null hypothesis when it is actually true
Holm-Bonferroni method	A more powerful alternative to the traditional Bonferroni correction to adjust the p-values and control the FWER
Box plot	Graphical representation to summarize the distribution of a sample: median, quartiles, and outliers
Heatmap	Graphical representation: individual values are displayed in a matrix format, with colors representing magnitudes
Odds	The ratio of the probability of an event occurring to the probability of it not occurring
Odds Ratio (OR)	The ratio of the odds of an event occurring in one group compared to another group
Priori probability	Probability of an event or outcome occurring before observing any data or experimental result
Posteriori probability	Probability of an event or outcome occurring given known information or evidence
Sensitivity analysis	A technique to study how the variation in the output of a model can be attributed to variations in its inputs

The COVID-19 pandemic, which began in December 2019[1], has had a profound global impact, resulting in millions of cases and fatalities worldwide. Although vaccines have slowed the rate of new infections, ongoing concerns about potential new waves and other health crises persist. In this context, advanced predictive models have become essential for improving clinical decision-making and patient care, particularly regarding hospital admissions for contagious diseases. Accurate risk assessment and timely interventions are crucial due to the rapid progression and variable severity of COVID-19. Advanced predictive models are needed to evaluate risks and optimize interventions based on individual patient profiles, ultimately improving patient outcomes and resource allocation.

Our study addresses these needs by introducing a novel Machine Learning (ML) methodology to construct predictive models specifically for COVID-19 patients for assessing ICU admission and death risks for COVID-19 patients admitted to the hospital. This model, which is a classifier, categorizes data points into predefined classes (icu, exitus and discharge) based on features such as demographics, vital signs, symptoms, comorbidities, and previous treatments available at hospital admission. The focus is on assessing the impact of therapeutic limits on ICU admission and death risks. To achieve this, we use a dataset comprising 3,362 SARS-CoV-2 infected patients admitted to hospitals in the south metropolitan area of Barcelona (Catalonia, Spain) between March and April 2020[2]. This dataset includes information on therapeutic limits, such as Non-Rebreather Mask (NRB) and Non-Invasive Mechanical Ventilation (NIMV). Therapeutic limits reflect critical decisions about the extent of medical interventions for COVID-19 patients based on factors such as age, physical activity, weight loss, and fatigue, which determine a patient's frailty level[3]. This inclusion allows to introduce ML methodology to address the impact of therapeutic limits on risk assessment for COVID-19 patients, which is a critical aspect of predictive modelling in healthcare. We evaluate the impact of therapeutic limits by dividing the patient cohort into two subsets: those with assigned therapeutic limits (NRB or NIMV) and those without, and constructing the ML predictive models for each data subset. Moreover, since the original number of variables is very high, we implement a feature selection process to manage the complexity of our large-scale dataset.

A key challenge in predictive modelling for COVID-19 is class imbalance within the dataset. For instance, the classes icu (10%), exitus (17%), and discharge (73%) are significantly imbalanced, which is exacerbated when partitioning the dataset based on therapeutic limits, as detailed in Table 1. This imbalance can lead to model bias, where the classifier performs better on majority classes and poorly on minority classes. Some researchers have addressed multiclass imbalance by merging minority classes[4], but this approach prevent for separate analysis of these classes.

To address this issue, our study employs a cost-sensitive classification approach. Traditional evaluation metrics like accuracy can be misleading in such scenarios, known as the *accuracy's paradox*. Focusing solely on

	discharge	exitus	icu	Total
Therapeutic limit	796 (60.58%)	483 (36.76%)	35 (2.66%)	1314
No-therapeutic limit	1657 (80.91%)	79 (3.86%)	312 (15.23%)	2048
Whole dataset	2453 (72.96%)	562 (16.72%)	347 (10.32%)	3362

Table 1. Distribution of the "event" output variable within each data subset and across the entire dataset.

minimizing the error rate (or maximizing accuracy) during classifier learning and validation is not appropriate, as it assumes equal severity in classification errors, which does not make sense in this scenario. Our approach assigns different weights to classification errors based on their consequences, improving model performance on imbalanced datasets.

Cost-sensitive classification methods can be implemented at three levels: data, algorithm, and output. **Data-level methods** encompass techniques such as oversampling, relabelling, and instance weighting, before training a cost-insensitive classification algorithm. However, these methods have several drawbacks: time-consuming and computationally expensive; risk of overfitting, particularly with oversampling; inefficient handling of class overlap, which can result in poor discrimination between classes; and lack of adaptability to changes in class distributions or the introduction of new data inputs. **Algorithm-level methods** involve modifying the original learning algorithm to account for misclassification costs. Despite their potential, they also face several challenges: increased algorithm complexity in modifying it; limited flexibility, as this modifications are specific to each algorithm, reducing their general applicability; difficulties in fine-tuning to balance misclassification costs effectively.

In contrast, **output-level techniques** function as wrappers or meta-algorithms, improving classification algorithms by adjusting thresholds –a process commonly referred to as *thresholding*. These techniques focus on post-processing the classifier's output to transform any cost-insensitive probabilistic classifier into a cost-sensitive one without altering the underlying algorithm. The advantages of these methods include: simplicity and flexibility, allowing they to be applied to any probabilistic classifier without requiring changes to the data or the underlying algorithm, thereby reducing the risk of overfitting; easy dynamic adaptation to variations in misclassification costs; transparency and interpretability, which are crucial in many fields, particularly in healthcare; scalability and computational efficiency, as they do not require extensive preprocessing or algorithm modifications.

Traditional *thresholding* methods, typically used for binary classification, fall short in multiclass scenarios. Overall, while *thresholding* can be a powerful tool for cost-sensitive classification, its application in multiclass scenarios is fraught with challenges that require careful consideration and often bespoke solutions. Some of these challenges are:

- Each class requires its own specific threshold based on the misclassification costs associated with it, but handling multiple thresholds at the same time is more complex. In addition, with more thresholds to tune, there is a greater risk of overfitting.
- Determine the appropriate thresholds for each class to minimize the overall misclassification cost is intricate, as it involves balancing the trade-offs between different misclassification error rates and costs.
- The threshold for one class can affect the performance of the other classes due to the inter-class dependencies, thus complicating the optimization process.
- Evaluating the performance is more complex due tot the variety of possible misclassifications. This requires more sophisticated evaluation metrics and analysis to accurately measure the impact of thresholding.
- Finding the optimal thresholds for each class can be computationally expensive.
- Interpreting the results of a thresholding approach in a multiclass setting can be challenging, making difficult for practitioners to understand and trust the model's predictions, especially in critical applications like healthcare.Our study introduces the *Multi-Thresholding meta-algorithm* (MTh), which dynamically adjusts label probabilities based on factors derived from the probability distribution and misclassification costs. Any instance is then classified with the label having the highest adjusted probability (following the MAP criterion), offering a more nuanced approach to handling minority classes without merging them. Figure 1 schematically illustrates this process. See Section 3.5 for detailed information on MTh and Algorithm 1 for pseudocode. Additionally, we discuss some key properties in Appendix D.

The MTh meta-algorithm enjoy the benefits of the **output-level techniques**, and in particular integrates with any probabilistic classifier, such as Naive Bayes (NB) or Support Vector Machines (SVM). In our experiments,

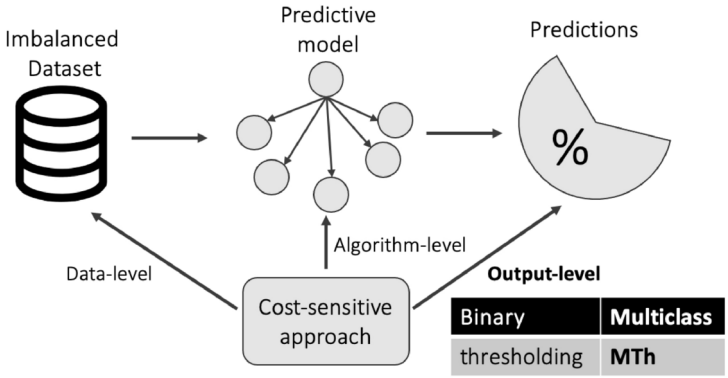


Fig. 1. Scheme illustrating the three levels (data, algorithm and output) at which the cost-sensitive approach can be applied in classification.

we consider integrating MTh with various state-of-the-art models, including Neural Networks (NN), Support Vector Machine (SVM), and Random Forest (RF). This represents a significant advancement over traditional methods by extending thresholding from binary to multiclass classification settings.

On the other hand MTh does not suffer from the disadvantages of traditional thresholding methods when applied to multiclass classification that we have just discussed as it avoids the need to determine thresholds, which is the intrinsic source of the method's issues. Furthermore, in this context, the evaluation metric naturally associated with the cost-sensitive approach, Total Cost (TC), is perfectly appropriate and useful.

Our study also constructs different types of Bayesian Networks (BN) as *explanatory* or *white-box* models to elucidate variable interdependencies and the rationale behind risk assessments. We explore how these dependencies differ between patients with and without therapeutic limits, aiming to enhance more informed patient care decisions. It is worth noting that both the predictive and the explanatory models for the subpopulations of patients with and without therapeutic limits may, and indeed do, exhibit differences.

In summary, the key contributions of our research are as follows:

- Novelty of the MTh meta-algorithm: We introduce the MTh meta-algorithm, which extends traditional thresholding from binary to multiclass classification. MTh fills a critical gap in output-level methodologies for multiclass scenarios by employing a cost-sensitive approach to handle imbalanced datasets (see Figure 1). Unlike conventional algorithms that rely on static or less flexible thresholds, MTh dynamically adapts to different data subsets and scenarios, leading to more precise and finely-tuned predictions.
- Resolution of traditional thresholding issues: MTh circumvents the common issues associated with traditional thresholding methods applied to multiclass classification, such as the complex task of setting appropriate thresholds. By eliminating the need for these thresholds, MTh effectively addresses these challenges and utilizes the Total Cost metric, which aligns naturally with cost-sensitive classification. This approach is particularly well-suited for evaluating and optimizing predictions in imbalanced datasets.
- Application and impact: Our model effectively assesses ICU admission and death risks for COVID-19 patients, taking into account therapeutic limits. This capability provides valuable insights that enhance clinical decision-making and resource allocation, potentially improving patient outcomes.
- Integration and adaptability: The MTh meta-algorithm can be seamlessly integrated with any probabilistic classifier, including state-of-the-art models. This compatibility represents a significant advancement over traditional methods that may lack such flexibility, thereby enhancing the overall predictive modelling capabilities.
- Addressing class imbalance: By employing a cost-sensitive classification approach, our research tackles the issue of class imbalance, which is crucial for accurate risk prediction. This is reflected in the application of the MTh meta-algorithm for multiclass classification and the use of the Total Cost (TC) metric, both of which account for misclassification costs effectively.These contribution collectively advance predictive modelling in healthcare, offering valuable tools for managing patient risks and optimizing interventions.

The paper is structured as follows: Section 2 reviews related research. Section 3 details the materials and methods, including the dataset and preprocessing in Sections 3.1 and 3.2, respectively. Section 3.3 introduces Bayesian Networks used in our models. Section 3.4 covers validation and evaluation metrics. Section 3.5 presents the MTh meta-algorithm and its implementation, discussed further in Section 3.6. Results are presented in Section 4, followed by practical examples in Section 5. The paper concludes with final remarks in Section 6. Appendices A), B and C provide additional tables and figures, and Appendix D offers theoretical justification for the MTh meta-algorithm.

Literature review

In this section, we provide a concise overview of relevant research related to the present study, covering various aspects of the topic. Overall, the intersection of COVID-19 research and ML techniques highlights a growing trend towards using advanced computational methods to enhance predictive capabilities and improve patient outcomes. This integrated approach underscores the evolving landscape of medical research, where traditional statistical methods and modern ML techniques complement each other to address complex health challenges.

Numerous research studies have explored COVID-19 using various statistical and ML methodologies to understand and predict different aspects of the disease. Early studies primarily focused on epidemiological aspects and potential risk factors. For instance, research has examined significant associations such as the link between blood type and disease severity[5], and the connection between myocardial injury and disease prognosis among hospitalized patients[6]. Some studies have gone further by constructing statistical models to predict disease risk. For example[7], uses a multivariate logistic regression model to predict the risk of death within 30 days for COVID-19 patients in emergency rooms. Similarly[8], used Cox regression analyses to identify factors associated with mortality in hospitalized COVID-19 patients, marking a significant advancement in pinpointing clinical and laboratory predictors of death.

The advent of ML has led to the development of advanced predictive models that go beyond traditional statistical approaches. ML techniques have enhanced various aspects of medical research and practice, becoming crucial in predicting patient outcomes and improving diagnostic accuracy in medicine. For instance, ML models have streamlined data analysis in intensive care units (ICUs), aiding in sepsis prediction and improving patient care allocation[9]. Other applications include predicting survival in heart failure patients based on serum creatinine and ejection fraction[10]. We are interested in the integration of ML into COVID-19 research for medical diagnostic. For example, deep convolutional neural networks have been used to detect COVID-19 from chest X-ray images[11], and more recently deep learning has been applied to identifying COVID-19 patients from computed tomography scans[12]. Given their reputation as effective classification algorithms in various fields and

their *white-box* nature, which differentiates them from other state-of-the-art supervised ML techniques like neural networks, we focus on Bayesian Networks (BNs).

BNs represent probabilistic relationships among variables, combining principles from graph theory, probability theory, computer science, and statistics[13]. They are versatile models of supervised learning that transparently illustrate the relationships between variables involved in a phenomenon, enabling both effective predictions and the generation of valuable knowledge. BNs have proved effective in various fields, including document classification, image processing, spam filters, speech recognition, robotics, semantic search, and operational risk assessment[14]. They have even found application in criminal profiling, such as identifying forest arsonists[15]or multi-victim homicides[16]. In the medical field, BNs have been successfully used to predict outcomes in areas such as pancreatic cancer[17]and ICU patients[18,19]. During the COVID-19 pandemic, BNs demonstrated their utility in predicting infection likelihood and disease severity using mobile applications powered by Bluetooth technology[20]. They have also been applied to predict COVID-19 outcomes based on symptoms[21], assess infection risk by ethnicity or religion[22], and predict qRT-PCR results[23]. Furthermore, BNs have contributed to predicting COVID-19 pneumonia outcomes from chest computed tomography scans evaluated by independent radiologists[24].

While many techniques exist for handling binary classification problems, there is a lack of satisfactory solutions for multiclass classification in imbalanced datasets, both within and outside the context of cost-sensitive learning. Cost-sensitive learning, in general, addresses the challenge of classification when different misclassification costs are involved, which is a common scenario in medical datasets. Some works have considered this issue, by assigning weights to training examples, which is not very effective, for example, if some classes have significant overlap in the feature space. For instance[25], introduces a method for multiclass classification within this framework that employs an iterative scheme for example weighting combined with a binary classification algorithm. Meanwhile[26], discusses the *rescaling* approach, which aims to rebalance classes according to their costs by assigning weights to the training examples based on their class. Although this method proves to be effective in the binary case, it often falls short in the multiclass setting. In such situations, the authors recommend decomposing the multiclass classification problem into a series of two-class problems to achieve better results.

Handling imbalanced medical data in the field of cost-sensitive classification is a rapidly evolving research area in ML[27,28]. At the data-level, SMOTE (Synthetic Minority Over-sampling TEchnique)[29]have become a widely adopted oversampling method in medical diagnostics. Innovations include combining SMOTE with Tomek links to balance medical data[30], integrating SMOTE with edited nearest neighbor[31], and applying SMOTE along with modified particle swarm optimization[32]. Alternative oversampling methods, such as BOSME (Bayesian network-based Over-Sampling MEthod), address SMOTE's limitations by accommodating non-continuous data[33].

Algorithm-level approaches for cost-sensitive classification, which involve modifying classification algorithms themselves, are less common but ther are some notable exceptions. These include cost-sensitive decision trees integrating game theory principles[34], incorporating feature ranking capabilities into a cost sensitive ensemble for classifying chronic kidney disease[35], and cost-sensitive variants of XGBoost applied to datasets related to breast cancer detection[36]. Additionally[37], provides a comprehensive list of cost-sensitive learning algorithms that modify loss functions to prioritize minority classes.

Finally, output-level adjustments, such as various binary *thresholding* methods, have been explored to improve cost-sensitive classification performance. Examples include *MetaCost*([38]), which employs bagging in decision trees to generate accurate probability estimates, CostSensitiveClassifier ([39]), *Cost-sensitive Naive Bayes*[40], and *Empirical Thresholding*([28]), which focuses on improving probability estimate calibration.

Materials and methods

Dataset overview

The dataset utilized in this study was generously provided by the Bellvitge Biomedical Research Institute (IDIBELL). Due to the retrospective nature of this study and the use of anonymized data, informed consent was waived as authorized by the Ethics Committee of Bellvitge University Hospital. This committee ensures compliance with national data protection legislation, including Spain's Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD) (BOE number 294, December 6, 2018, pages 119788 to 119857), and the European Union's General Data Protection Regulation (Regulation EU 2016/679). The methodologies employed for data processing in this study, aimed at scientific research and statistical analysis, fully adhere to these regulations, ensuring the legality and protection of patient data.

The dataset encompasses a cohort of 3,362 patients admitted to five hospitals or hospital consortia in the southern metropolitan area of Barcelona (Catalonia, Spain) with confirmed SARS-CoV-2 infection. These healthcare institutions include the Consorci Sanitari de l'Alt Penedès i Garraf (402 cases), Consorci Sanitari Integral (895 cases), Parc Sanitari Sant Joan de Déu (538 cases), Hospital de Viladecans (404 cases), and Hospital Universitari de Bellvitge (1,123 cases). The patients were admitted between March 1, 2020, and April 30, 2020.

For each patient, we compiled a comprehensive set of characteristics, including demographic variables, comorbidities, and treatments administered since hospital admission. Initially, we dealt with a total of 1,012 variables (1,007 features and 5 output variables) distributed across 10 different files, which required extensive cleaning and preprocessing. This lead to a "large-scale" classification challenge due to the significant number of features, which can pose computational hurdles. To address this complexity, we applied a specific feature selection approach during the preprocessing phase (see Section 3.2 for more details). Regarding the output variables, we merged several into a single output variable called event. This variable signifies the patient's outcome and encompasses the values of discharge/icu/exitus, indicating whether the patient was discharged from the hospital without entering the ICU, admitted to the ICU, or passed away in the hospital without ICU admission.

Of noteworthy significance is the development of two distinct predictive models: one for patients with therapeutic limits and another for those without. This segregation is prompted by the expected differences in behaviour between these two patients subgroups, primarily due to their unique characteristics. One notable difference is the variation in the minority class labels for the output variable event. Among patients with therapeutic limits, the minority class label is icu (accounting for 2.66%), while for the remaining patients, it is exitus (representing 3.86%), as depicted in Table 1. A primary objective of this study is to compare these two patient populations, identifying distinctions both in terms of primary risk factors and their corresponding predictive models.

Preprocessing

First, we adopted an expert-driven preprocessing approach to refine the dataset by eliminating redundant and irrelevant variables. Specifically, we excluded variables related to post-admission patient health status and information recorded before admission, as these were deemed nonessential for our research by medical experts. To streamline the dataset's efficiency and information content, we also merge related variables.

To facilitate the use of standard Bayesian Networks, it is imperative for all variables in the model to be of the *factor* type, which encompasses binary, categorical or discrete variables with a finite number of possible values. Consequently, we discretized continuous variables c-reactive protein, d-dimer, lactate, and lymphocytes into intervals based on the equal-frequency criterion, with slight adjustments to create more memorable or rounded intervals, provided these adjustments do not significantly imbalance them. This adjustment method maintains the integrity of the data distribution while making the intervals more intuitive and user-friendly. However, for variables age and O_2 saturation, we determined categories based on domain expertise, following the guidelines provided by medical specialists to ensure the most relevant information was captured. Specifically, for O_2 saturation, which is a percentage, the categorization is as follows: < 90 as "hypoxia", $[90,\ 95)$ as "low", and ≥ 95 as "normal". The values of the discrete variable Charlson index[41], which can take a finite but large number of possible values, were grouped into 5 categories using a binning method. The criterion for this binning was to ensure that the frequencies among the categories are relatively balanced. This approach helps to maintain a more even distribution across the categories, facilitating more robust analysis and interpretation.

Addressing missing data was another crucial aspect. Some variables within our dataset contained missing values, typically due to specific data unavailability for various reasons determined by the hospital's medical team. In response, we opted to remove certain variables with a substantial amount of missing data. For the remaining variables, instead of opting for imputation, we introduced a dummy category labeled "unknown" to account for missing values.

Considering the aforementioned points, after the initial feature selection phase, we streamline the number of input variables designated for constructing the predictive model. This not only reduces computational complexity but can also improve model performance in some cases. Traditional statistical-based feature selection methods involve assessing the relationship between each input variable and the output, selecting variables with the strongest associations. However, these methods rely on choosing an appropriate statistical measure to quantify the strength of these relationships, which can be challenging for any given dataset. Instead, we adopted the approach detailed in[42], which centers around the concept of the *Markov blanket* in a Bayesian Network. This approach has demonstrated remarkable efficiency in handling high-dimensional data, enabling the development of a sparse classifier that employs only a subset of the most informative features. This effectively reduces the problem's complexity and can lead to improved performance in both training and prediction, making it particularly advantageous for large-scale datasets. For a comprehensive breakdown of the characteristic set following this preprocessing, specific to each of the two data subsets (patients with and without therapeutic limits), please see Appendix A.

Bayesian networks: a supervised ML tool

Bayesian Networks are probabilistic models that represent the relationships among variables influencing a particular phenomenon. For a given set of random variables, a Bayesian Network models their joint probability distribution. The graphical component of this model is a directed acyclic graph (DAG), where the nodes represent the random variables, and the directed arcs connecting these nodes indicate conditional dependencies (not necessarily causal) governed by the Markov condition. According to the Markov condition, each node in the DAG is independent of all other nodes that are not its descendants, given its parent nodes.

Bayesian inference involves updating the probabilities within the network based on observed evidence. This process requires calculating posterior probabilities using both the evidence and the prior probabilities. For predicting a query variable –in our case, the output variable "event"–, we employ the Maximum A Posteriori (MAP) criterion. This criterion selects the most probable instantiation of the event, with the corresponding probability termed the *confidence level* of the prediction.

In constructing both explanatory and predictive models for the variable event, we explore three types of Bayesian Networks: *Naive Bayes*, *Augmented Naive Bayes* and *Tree Augmented Naive*[43].

1. **Naive Bayes (NB)**: This is the simplest for of a Bayesian Network, assuming that all features are conditionally independent of each other given the value of the class variable "event". Although this independence assumption may not always be valid, Naive Bayes has demonstrated good predictive performance in many scenarios. The DAG structure in NB is fixed, with directed edges originating from the class variable and pointing to each feature. Since the structure is predefined, no algorithm is required to learn it; however, the parameters are learned from the data using Maximum Likelihood Estimation (MLE). For that, we use the function bn.fit from the R package **bnlearn**([44]).

2. **Augmented Naive Bayes (AN)**: This variant allows for additional directed edges between features, enabling both the structure (directed edges) and parameters to be learned from the data. The structure is determined using the function hc from the R package **bnlearn**, which implements a hill-climbing algorithm constrained by a whitelist that enforces directed edges from the class variable to the features. The Bayesian Information Criterion (BIC) score function is used to guide the structure learning process.
3. **Tree Augmented Naive (TAN)**: TAN belongs to the *One-Dependence Estimators* (ODEs) family of models. It retains the simplicity of Naive Bayes while allowing for one additional dependency per feature. In TAN, in addition to the directed edges from the class variable to the features, each feature can also have a directed edge from at most one other feature. The TAN structure is learned using the tree.bayes function from the R package **bnlearn**, with parameters consistently estimated using the MLE method.For Bayesian inference with these Bayesian Networks, we use the R package **gRain**[45]. Figure 2 provides examples of DAGs corresponding to these three types, illustrating a class variable C and five features, $X_1, \ldots, X_5$.

Validation and performance evaluation

To conduct the validation process and compare the different models, we employ a standard K-fold cross-validation with $K = 10$. This approach ensures that the entire dataset is used for both training and validation across all K iterations. The primary objective of cross-validation is to assess the model's ability to make predictions on new, unseen cases, thereby helping to detect issues such as overfitting. In this process, the dataset is randomly divided into K roughly equal-sized folds. In each iteration, one fold is set aside for validation while the remaining folds are used to train the model. Upon completing the process, we generate K *confusion matrices* for each classifier, from which we can compute the relevant performance metrics.

We denote the class labels as $\{c_1, \ldots, c_r\}$ and represent a general confusion matrix from the validation procedure as $(C_{ij})_{i,j=1,\ldots,r}$. Here, C_{ij} indicates the number of instances in the validation dataset that truly belong to class c_j but are predicted by the classifier as class c_i. The total number of instances in the validation dataset is given by $N = \sum_{i=1}^{r} \sum_{j=1}^{r} C_{ij}$. While binary classification ($r = 2$) is most common, our class variable event comprises $r = 3$ categories (discharge/exitus/icu), so we focus on the multiclass scenario.

In cost-insensitive classification, evaluation metrics do not account for the varying importance of different classification errors. The most common metric is *accuracy*, calculated as the proportion of correctly classified instances (i.e., *accuracy* $= \sum_{i=1}^{r} C_{ii}/N$). The *error rate* is the complement of *accuracy*, defined as the proportion of misclassified instances, i.e., $error\ rate = 1 - accuracy$.

In classification tasks, particularly those involving imbalanced datasets, traditional *accuracy* may not be an adequate measure of model performance due to the risk of bias toward the majority class, a situation known as the *accuracy paradox*. Relying solely on *accuracy* (or *error rate*) can lead to selecting a poorly performing model. To address this issue, *Balanced Accuracy* (BA) is introduced. BA provides a more balanced evaluation by considering the performance across all classes equally, regardless of their proportions in the dataset. It is defined as the average of the true positive rates (TPR) for each class, offering a more comprehensive view of the model's ability to correctly classify instances from each class. This metric is particularly useful in cases where some classes are significantly underrepresented, ensuring that the model's performance is not overly influenced by the majority class. Mathematically, BA is expressed as:

$$\mathrm{BA} = \frac{1}{r} \sum_{j=1}^{r} \frac{C_{jj}}{n_j}$$

where $n_j = \sum_{i=1}^{r} C_{ij}$ is the total number of cases belonging to class c_j.

While BA provides a valuable metric by considering all classes equally, it does not account for the different consequences of misclassification errors. In many real-world applications, the cost associated with different types of errors varies significantly. This is particularly true in our case, where misclassifying a critically ill patient as stable could have far more severe implications than the reverse error. To address this, we introduce a *cost-sensitive* metric, Total Cost (TC), which incorporates the specific costs associated with different types of

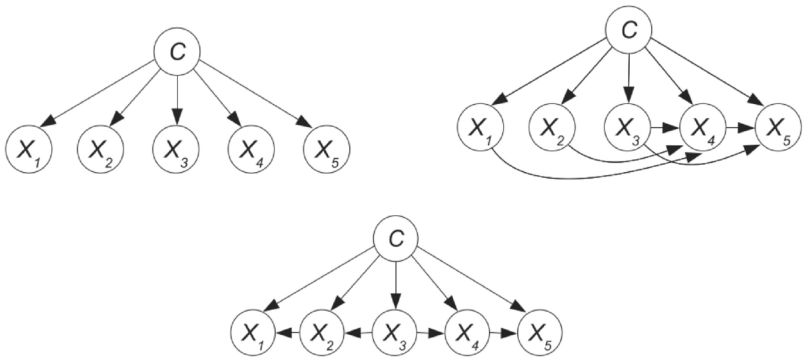


Figure 2. Left: Naive Bayes (NB). Center: Augmented Naive (AN). Right: Tree Augmented Naive (TAN).

misclassification. Using a predefined cost matrix, TC quantifies the overall cost of errors made by the classifier, reflecting the practical importance of each type of misclassification. This approach aligns the model evaluation with real-world priorities and consequences. Mathematically, TC is calculated by summing the products of the number of misclassifications and their corresponding costs.

The cost of classifying a patient into class c_i when they actually belong to class c_j is denoted as m_{ij}, where $m_{jj} = 0$ signifies no error and no cost. The **Total Cost (TC)** is then calculated as:

$$\mathrm{TC} = \frac{1}{N} \sum_{i,j=1}^{r} C_{ij}\, m_{ij}\,.$$

Here, $(m_{ij})_{i,j=1}^{r}$ represents the *cost matrix*. In our study, medical experts assigned a minimum cost of 1 for misclassifying a patient that actually belong to discharge as icu or exitus. The maximum cost, set to $\alpha = 10$, corresponds to the most critical error of misclassifying exitus as discharge, with a lesser but still significant cost for misclassifying as icu. The cost of misclassifying a patient who should be in the ICU is $\beta = 8$ when incorrectly classified as discharge, and half that when classified as exitus. Thus, the cost matrix is defined with $\alpha = 10$ and $\beta = 8$ in equation (1).

$$\begin{array}{cc} & \text{observed} \\ & \begin{array}{ccc} \texttt{discharge} & \texttt{exitus} & \texttt{icu} \end{array} \\ \text{predicted} \begin{array}{r} \texttt{discharge} \\ \texttt{exitus} \\ \texttt{icu} \end{array} & \begin{pmatrix} m_{11}=0 & m_{12}=\alpha & m_{13}=\beta \\ m_{21}=1 & m_{22}=0 & m_{23}=\frac{\beta}{2} \\ m_{31}=1 & m_{32}=\frac{\alpha}{2} & m_{33}=0 \end{pmatrix} \end{array} \tag{1}$$

We assume the same cost matrix for both patient subsets, with and without therapeutic limits, though specific cost matrices could be defined for each subset if necessary. The Total Cost is calculated as follows:

$$\mathrm{TC} = \frac{1}{N}\left(C_{12}\,\alpha + C_{13}\,\beta + C_{21} + C_{23}\,\frac{\beta}{2} + C_{31} + C_{32}\,\frac{\alpha}{2}\right).$$

In the experimental phase, we perform a parameter sweep for α and β. We systematically vary these parameters to observe their impact on the predictive models using both the TC and the BA metrics. The parameters ranges are chosen to include $\alpha = 10$ and $\beta = 8$, with the constraint that β must be less than α. Specifically, α ranges from 2 to 20, and β ranges from 2 to α, with both parameters being integers.

When $\alpha = \beta = 2$, the costs of misclassifying a patient as exitus when the true class is icu, and vice versa, is set to 1. The same as for misclassifying a discharged patient as exitus or icu. This scenario reflects a situation where all errors are treated as equally impactful, except for misclassifications involving discharge when the true class is exitus or icu, which are assigned double the cost due to their greater impact.

The constraint $\beta \leq \alpha$ highlights that the most critical misclassification errors are those where the true condition exitus but the patient is classified otherwise. We employ a grid search approach to explore the 190 possible pairs $(\alpha,\ \beta)$. This analysis helps us understand the sensitivity of our results to these parameters, identify trends or patterns, and evaluate how these parameters influence model performance.

Thresholding: an indirect cost-sensitive meta-learning approach

As previously mentioned, our dataset is imbalanced, with two minority classes (icu and exitus) and one majority class (discharge). This highlights the need for a specific approach. We use a cost-sensitive learning approach, the *thresholding* method, which we term the *Multi-Thresholding meta-algorithm* (MTh). This is an indirect cost-sensitive approach that acts as a wrapper, transforming any cost-insensitive probabilistic classifier into a cost-sensitive one. This transformation is achieved by post-processing the classifier's output, which includes the probability distribution assigned to the classes. MTh involves adjusting this probability distribution based on the expected costs associated with misclassification, ensuring that the class with the highest adjusted probability is selected. These expected costs are calculated using the cost matrix described in equation (1), derived from expert knowledge. For an overview of the algorithm, refer to Algorithm 1 below.

In our specific application, if a classifier assigns a probability distribution $p = (p_1,\ p_2,\ p_3)$ to the class labels c_1 :discharge, c_2:exitus, and c_3:icu, the adjusted probabilities are calculated by dividing p_i by ω_i. Here, ω_i represents the expected cost associated with misclassifying an instance when assigning it label c_i. The expected costs are calculated as follows:

$$\begin{aligned} \omega_1 &= m_{12}\,p_2 + m_{13}\,p_3 = \alpha\,p_2 + \beta\,p_3, \\ \omega_2 &= m_{21}\,p_1 + m_{23}\,p_3 = p_1 + \frac{\beta}{2}\,p_3, \\ \omega_3 &= m_{31}\,p_1 + m_{32}\,p_2 = p_1 + \frac{\alpha}{2}\,p_2\,. \end{aligned}$$

The adjusted probabilities are then computed as: $\widetilde{p}_1 = \dfrac{p_1}{\omega_1}\,,\quad \widetilde{p}_2 = \dfrac{p_2}{\omega_2}\,,\quad \widetilde{p}_3 = \dfrac{p_3}{\omega_3}\,.$

Strictly speaking, we should normalize these adjusted probabilities by dividing each $\widetilde{p}_i$ by the sum of all of them to ensure they form a probability distribution. However, this step is unnecessary for our purpose, as we only need to identify the class that maximizes the adjusted probabilities. To simplify, we refer to $\widetilde{p}_i$ as "probabilities",

although they are technically not. Notably, as the cost associated with misclassifying a label increases, the adjusted probability for that label decreases, making it less likely to be selected as the final prediction. This clearly illustrates the implementation of a cost-sensitive classification algorithm. While the results in Section 4 depend on the specific values in the cost matrix (1), the algorithm's procedure remains consistent regardless of these values. Appendix D further elaborates on and proves some properties of this algorithm.

Input labels $c_1, \ldots, c_r$, probability distribution assigned to the labels $(p_1, \ldots, p_r)$, cost matrix $(m_{ij})_{i,j=1,\ldots,r}$

Output The predicted class (label) c^*

1: **for** i in $1:r$ **do**

2: **compute** $\omega_i = \sum_{j=1}^{r} m_{ij}\, p_j$ (expected cost of misclassifying an instance as c_i)

3: **adjust** probabilities by $\widetilde{p}_i = \frac{p_i}{\omega_i}$

4: **end for**

5: $\ell = \arg \max_{i=1,\ldots,r} \widetilde{p}_i$

6: $c^* = c_\ell$

return c^*

(If the maximum adjusted probability is not unique, a tiebreaker rule is used to select one of the labels.)

Algorithm 1. The Multi-Thresholding meta-algorithm, MTh

When the MTh meta-algorithm is not applied, meaning we take $\omega_i = 1$ for all *i*, the adjusted probabilities $\widetilde{p}_i$ become identical to the original probabilities p_i. In this case, the classification follows the standard Maximum A Posteriori (MAP) criterion for multi-class classification. Thus, comparing the use of MTh with not using it is effectively a comparison between the MTh and the MAP criterion, the latter being the reference standard for multi-class classification.

Implementation of the experimental phase

All computational aspects of model implementation were conducted using the R programming language[46]. For each of the two data subsets –one for patients with therapeutic limits and the other for patients without such limits– we constructed three types of Bayesian Networks as predictive (and explanatory) models: *Naive Bayes* (NB), *Augmented Naive* (AN) and *Tree Augmented Naive* (TAN), both with and without applying the MTh meta-algorithm.

Additionally, to provide a comprehensive comparison and assess how our proposed approach compares with other state-of-the-art predictive models, we included three advances ML models: *Neural Network* (NN), *Support Vector Machine* (SVM), and *Random Forest* (RF), the latter being constructed as an ensemble of 100 *decision trees*. We constructed these models using the mlNnet, mlSvm and mlRforest functions from the R package **mlearning**.(Authors: Ph. Grosjean & K. Denis. https://doi.org/10.32614/CRAN.package.mlearning).

Our primary objective was to experimentally validate the use of the MTh meta-algorithm for multi-class classification with imbalanced data. Once this validation was achieved, we focused on selecting the model with the best predictive performance from those tested. The comparison between models (using the MTh meta-algorithm) was structured as follows: First, we evaluated the three Bayesian Network models against each other, and separately compared the other three ML models among themselves. After identifying the best-performing model within each group, we then compared these "winning" models to determine the overall best performer. This comparison was conducted separately for each subpopulation –patients with and without therapeutic limits– and for each performance metric used: Total Cost (TC) and Balanced Accuracy (BA).

Although we also included the Balanced Accuracy metric, our main reference is the TC metric, which is specifically designed to account for the varying weights assigned to classification errors, making it particularly suited to our cost-sensitive approach. The BA metric, while valuable for its balanced evaluation of class performance, does not account for the different consequences of misclassification errors. Therefore, while the TC metric aligns with our cost-sensitive methodology, BA provides a complementary view to ensure a comprehensive evaluation of model performance in handling imbalanced data.

The *K*-fold cross-validation procedure, with *K* set to 10, resulted in each model producing 10 confusion matrices. This provided a sample of 10 values for both the TC metric and the BA metric (as detailed in Section 3.4). To compare the models, we conducted appropriate statistical hypothesis tests. Initially, we used the Shapiro-Wilk[47]test to assess data normality. Based on the results, we applied either the paired Student's t-test[48]or the Wilcoxon signed-rank test[49] to compare pairs of samples, depending on whether the data could be assumed to follow a normal distribution. To account for multiple comparisons between models, we adjusted p-values using the Holm-Bonferroni method (We opted for paired tests with adjusted p-values over ANOVA with multiple comparisons due to the nature of our context. Paired tests are specifically designed for comparing paired samples, which aligns with the cross-validation setup. Furthermore, paired tests offer clear and interpretable results for pairwise model comparisons, simplifying the process of identifying superior-performing models.

As usual, throughout the paper superscripts • indicates statistical significance at 10%, ∗ at 5%, ∗∗ at 1% and ∗∗∗ at 0.1%, for all p-values). When applying p-value adjustments using the Holm-Bonferroni method in the context of multiple model comparisons, a significant challenge arises as the number of models increases. The Holm-Bonferroni method controls the family-wise error rate, thereby reducing the likelihood of Type I errors (false positives) during multiple hypothesis testing. However, as the number of comparisons grows, the adjustment becomes increasingly conservative, making it harder to detect statistically significant differences between models and raising the risk of Type II errors (false negatives), where genuinely significant differences are overlooked. To address this issue, we divided the six models into two groups. We first compared models within each group, identified the best-performing model in each, and then compared the "winners" against each other. This approach reduces the number of comparisons at each stage, thus mitigating the risk of overly conservative p-value adjustments and enhancing our ability to detect meaningful differences between models.

In the following section, we will present the validation results and perform a comparative analysis of the different models.

Results

The explanatory model

Among the six models we compared, only the three types of Bayesian Networks (BN) –Naive Bayes (NB), Augmented Naive (AN), and Tree Augmented Naive (TAN)– can be considered explanatory models. Bayesian Networks are categorized as "white-box" models, meaning their internal processes are transparent and interpretable. In contrast to "black-box" models such as Neural Networks, Support Vector Machines, and Random Forests, which may offer high predictive power but lack interpretability, BNs allow us to understand and explain how the predictions are generated. The probabilistic relationships between variables in these networks can be visualized and interpreted, making them particularly valuable in contexts where insight into the decision-making process is as important as the predictions themselves. This interpretability is a significant advantage in applications where understanding the reasoning behind a model's prediction is crucial, such as in medical decision-making or other high-stakes domains.

We use the R[46] package **bnlearn** and its strength.plot function to visualize the Bayesian network structures, represented by the Directed Acyclic Graphs (DAGs) in Figures 6 and 7 in Appendix B. These DAGs illustrate the conditional independence relationships entailed by the AN and TAN models for the "therapeutic limit" data subset, similar to Figures 8 and 9 for the "no-therapeutic limit" data subset.

In these plots, the thickness of directed arcs reflects the strength of the dependencies they represent, with thicker lines indicating stronger relationships. To quantify the strength of each arc (or feature) while keeping the rest of the network structure fixed, we use the arc.strength function, which provides a p-value associated with the conditional independence test for removing the arc from the network. Smaller p-values indicate stronger relationships.

Tables 11 and 12 in Appendix B present the most influential features for predicting event, considering only p-values less than 0.05. Notably, among the two demographic features for the "no-therepeutic limit" patients, age shows a significantly stronger influence on predicting the output variable event. As expected, the limit type is highly influential for the "therapeutic limit" patients. Additionally, only three symptoms –confusional syndrome, dyspnoea and rhinorrhea– are influential for both data subsets. The only comorbidity that holds influence across both data subsets is dementia. Furthermore, only two previous treatments, acetylsalicylic acid and statins, play a notable role in risk prediction for both data subsets.

The predictive model: Thresholding vs no-thresholding

In the validation procedure, we conducted a comprehensive comparison of the Total Cost (TC) and Balanced Accuracy (BA) metrics across all classifiers (the three types of Bayesian Networks, NN, SVM and RF) both with and without the application of the *thresholding* meta-algorithm MTh. This comparison encompasses every integer value of α ranging from 2 to 20 and each integer value of β ranging from 2 to α, resulting in a total of 190 distinct comparisons.

Table 2 below summarizes the number of comparisons that significantly favour predictive models utilizing the MTh meta-algorithm (first multi-row) versus those without it (second multi-row) out of the total of 190 possible comparisons. Notably, an overall advantage is observed in favour of using the MTh meta-algorithm. We observe a similar pattern across the two data subsets with the TC metric, but a markedly different pattern when the BA metric is used. The majority of statistically significant results support the superiority of MTh with the TC metric, except in a few cases with small values of α and β. When considering the BA metric, the results are favourables to MTh for the data subset of patients without therapeutic limits. In the subset of patients with therapeutic limits, there are fewer statistically significant results. Some of these support the MTh meta-algorithm, corresponding to low values of α and high values of β ($\beta \leq \alpha$), while those that oppose its use correspond to high values of α and low values of β.

The experimental results strongly support the use of the MTh meta-algorithm when TC is used as performance metric, both for patients with and without therapeutic limits. When the BA metric is considered, the same trend is observed for the non-therapeutic limit data subset. However, for the "therapeutic limit" subset with the BA metric, the results are more balanced between favouring and opposing MTh.

Figure 3 illustrates the differences in the mean metric values between using and not using the MTh meta-algorithm (positive values favour MTh) across all predictive models, considering both patient groups. For simplicity in representing these differences in a line graph, we have reduced the parameter set by fixing $\alpha = 10$ and $\beta = 8$, as an example.

Based on these findings, we decided to adopt the use of the MTh meta-algorithm across both patient subsets and all predictive models. This experimental validation confirms the effectiveness of the MTh meta-algorithm,

	TC metric		BA metric	
	Therapeutic limit	No-therapeutic limit	Therapeutic limit	No-therapeutic limit
Favour MTh	NB 187	NB 171		NB 190
	AN 190	AN 155	AN 29	AN 190
	TAN 190	TAN 143	TAN 15	TAN 178
	NN 189	NN 188		NN 190
	SVM 190	SVM 190	SVM 4	SVM 190
	RF 189	RF 158	RF 8	RF 185
Against MTh			NB 35	
		AN 14		
		TAN 14		
			NN 50	
		RF 12	RF 7	

Table 2. Number of comparisons, out of 190 possible, favouring predictive models with or without the MTh meta-algorithm, based on the TC and BA metrics. Comparisons are conducted with $\alpha = 2, \ldots, 20$ and $\beta = 2, \ldots, \alpha$, and results are distinguished for the two data subsets: patients with and without therapeutic limits.

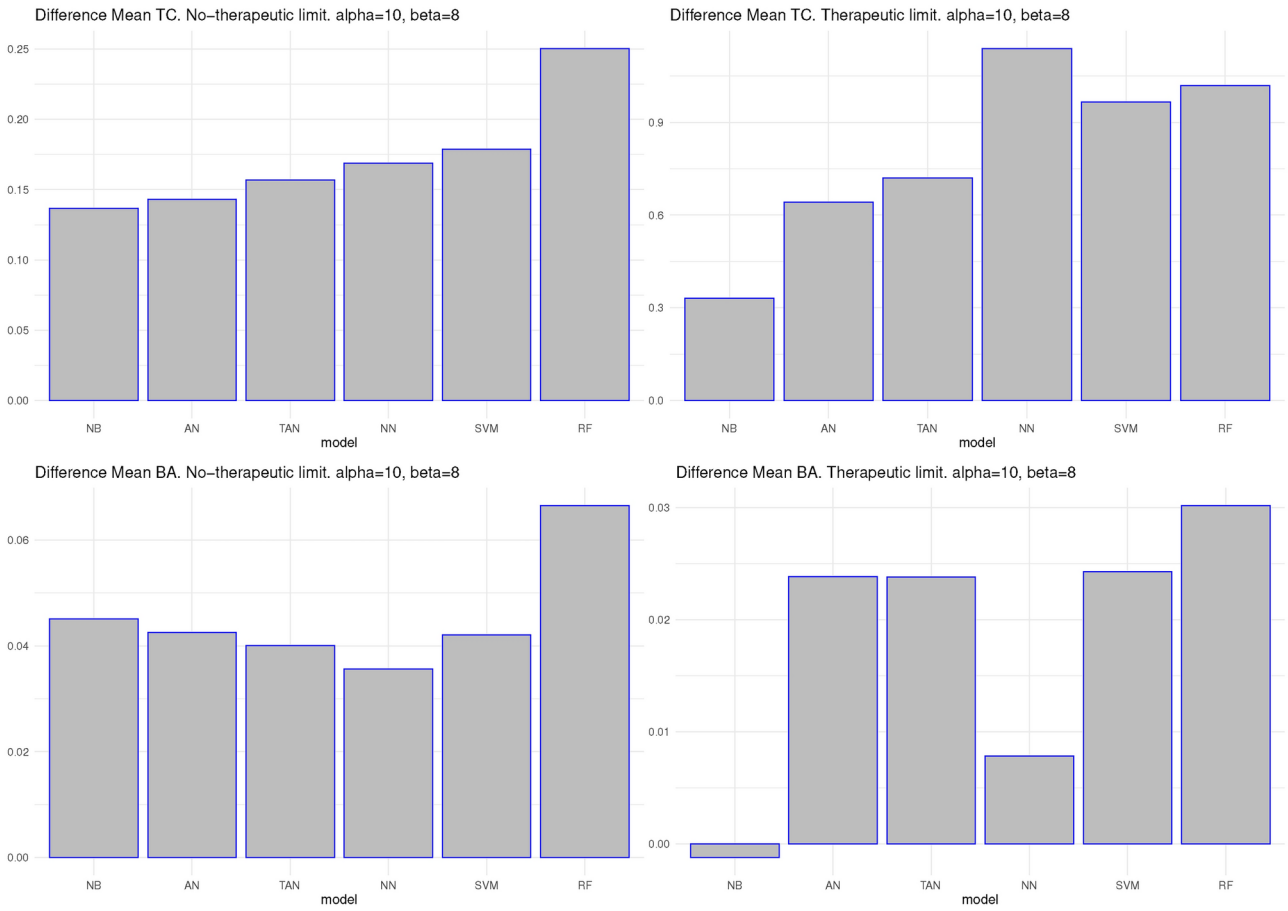


Figure 3. Bar plots illustrating the increase in the mean of the TC and BA metrics when applying the MTh meta-algorithm versus not using it, across the predictive models and the data subsets of patients with and without therapeutic limits. The plots are generated with parameters set to $\alpha = 10$ and $\beta = 8$. Positive values indicate that the MTh meta-algorithm improves both metrics.

particularly when using the TC metric, and also for the patients without therapeutic limits. In the following subsection, we will identify the preferred predictive model for each data subset and metric.

The predictive model: choosing the best

Using the *thresholding* meta-algorithm MTh, we compared the mean (or median) values of the two metrics –Total Cost (TC) and Balanced Accuracy (BA)– across different predictive models, including three types of Bayesian Networks (NB, AN and TAN) and three state-of-the-art models (NN, SVM and RF), for each parameter pair (α, β) where $\alpha = 2, \ldots, 20$ and $\beta = 2, \ldots, \alpha$. We used adjusted p-values for multiple comparisons, applying the Holm-Bonferroni method to ensure statistical rigor.

Table 3 displays the number of (α, β) pairs, out of the 190 possible, for which significant differences were found between models. This table highlights which model's mean (or median) is significantly higher (>) for each pair of models concerning the TC and BA metrics. Specifically, for each metric, it shows:

- The comparison among the three Bayesian Network classifiers (NB, AN, TAN) in the first multi-row.
- The comparison among the state-of-the-art models (NN, SVM, RF) in the second multi-row.
- A comparison between the best models from the two groups. They are NB and RF, respectively, with the TC metric, and with the BA metric for the "no-therapeutic limit" data subset. For the "therapeutic limit" data subset and the BA metric, NB is compared with SVM instead of RF. The results indicate a consistent advantage for NB over RF in the first case. Conversely, SVM shows slightly better performance for patients with therapeutic limits when using the BA metric.

Figure 4 presents box plots illustrating the distribution of TC values for $\alpha = 20$ and $\beta = 20$, providing a representative example of overall trends. These plots display results for the three models from both the "therapeutic limit" and the "no-therapeutic limit" data subsets, with the MTh meta-algorithm applied. Lower median TC values indicate better predictive performance, and the box plots visually confirm the superior performance of NB compared to the other models when $\alpha = \beta = 20$, as previously observed in Table 3. Similarly, higher median values indicate better predictive performance for the BA metric. The box plots corroborate the clear advantage of NB for patients without therapeutic limits and highlight the slight edge of SVM for patients with therapeutic limits, as previously noted in Table 3.

That is, after a rigorous model comparison, Naive Bayes (NB) consistently emerges as the preferred predictive model when evaluated with the Total Cost (TC) metric, and is particularly favoured for patients without therapeutic limits when using the Balanced Accuracy (BA) metric. Conversely, while Support Vector Machine (SVM) shows some preference as the predictive model for patients with therapeutic limits, this preference is less pronounced.

Figure 5 includes heatmaps that show how different combinations of parameters affect the mean metric values for the selected predictive model NB across both data subsets and both metrics, except for the data subset of patients with therapeutic limits and BA metric, for which the chosen model is SVM). The heatmaps depict the mean metric value as a function of $\alpha = 2, \ldots, 20$ and $\beta = 2, \ldots, \alpha$, with values derived from $K = 10$

TC metric		Therapeutic limit	No-therapeutic limit
NB vs. AN vs. TAN	AN > NB	187	162
	TAN > NB	180	190
	TAN > AN	1	131
NN vs. SVM vs. RF	RF > NN		4
	RF > SVM		10
	SVM > RF		117
NB vs. RF	RF > NB	9	190
BA metric			
NB vs. AN vs. TAN	AN < NB		124
	TAN < NB		187
	TAN < AN		67
NN vs. SVM vs. RF	NN < SVM	163	
	NN < RF	14	9
	SVM < RF		6
	RF < SVM	23	
NB vs. SVM (limit)	NB < SVM	36	
NB vs. RF (no limit)	RF < NB		190

Table 3. Number of pairs (α, β), out of the total 190 possible pairs, where one predictive model exhibits a significantly higher (>) mean (or median) metric compared to another, using the MTh meta-algorithm. The metrics considered are TC and BA. Only statistically significant results are reported; cells left blank indicate no significant difference.

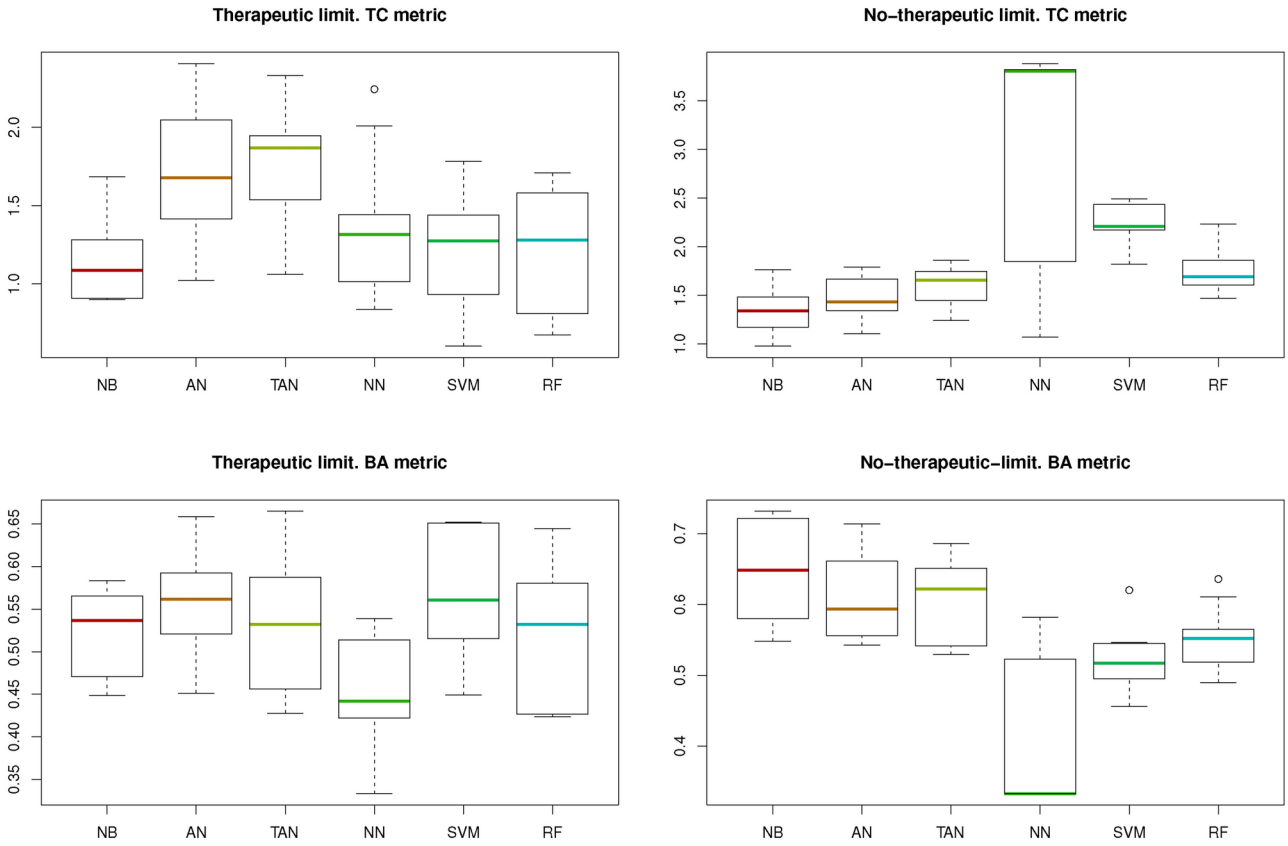


Figure 4. Box plots illustrating the TC and BA values obtained through cross-validation using the MTh meta-algorithm for both data subsets, with $\alpha = \beta = 20$. Lower TC values indicate better performance, whereas higher BA values reflect better predictive behaviour.

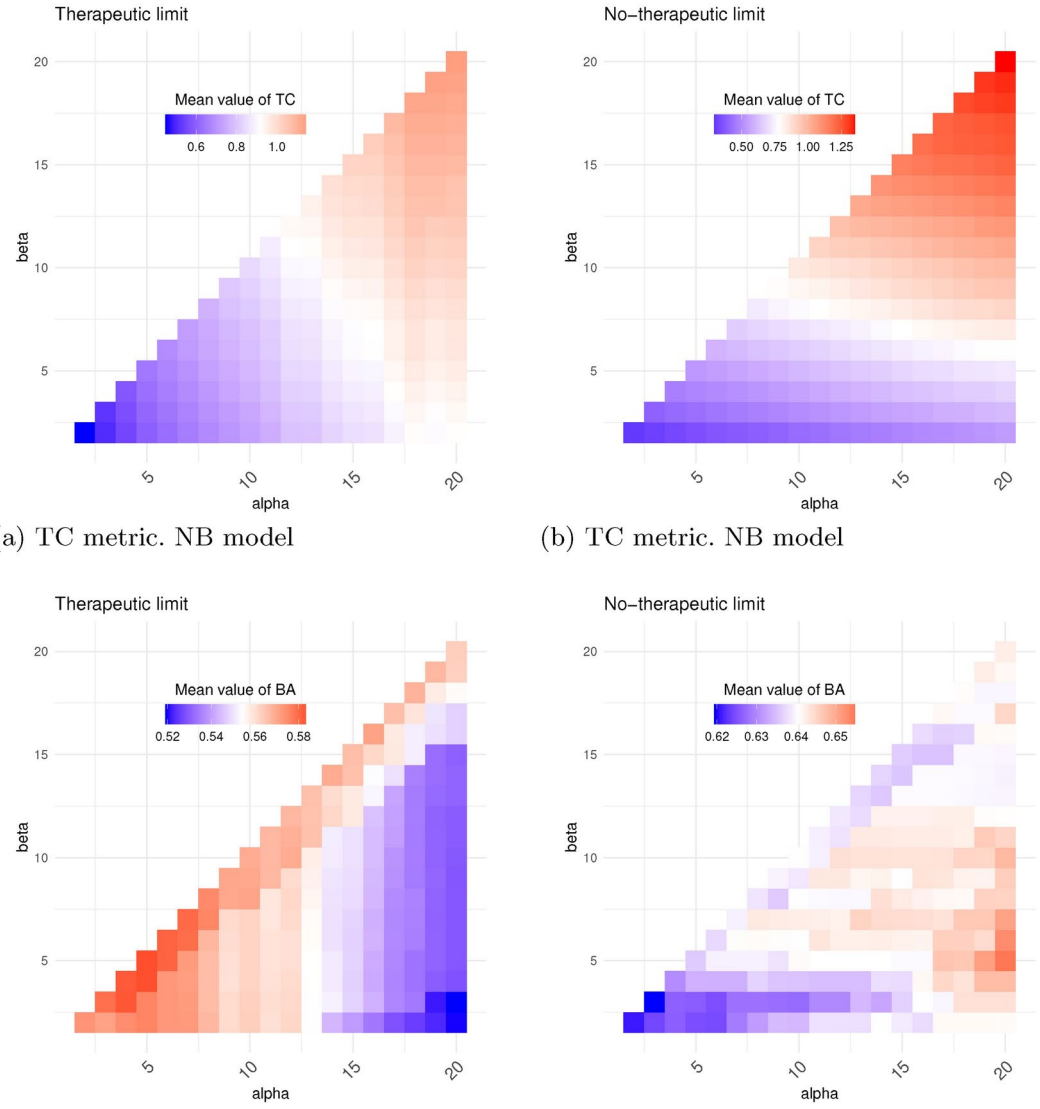
measurements per (α, β) pair during cross-validation. As anticipated, higher values of α or β generally result in increased mean TC values. Additionally, the mean TC values are consistently higher for the model applied to patients without therapeutic limits, reflecting diminished performance. Conversely, when using the BA metric, this trend becomes less predictable and more variable, specially for the patients without therapeutic limit.

Case studies in risk assessment using the predictive model

In this section, we illustrate the capabilities of our predictive model with specific examples. Consider a COVID-19 patient admitted to the hospital with a therapeutic limit. The initial (*a priori*) risk estimates are as follows: `exitus` : 36.76%, `icu` : 2.66%, and discharge: 60.58%. We will now evaluate the updated (*a posteriori*) risk predictions based on the patient's specific characteristics using our selected predictive model. To measure the association between a patient's characteristic (factor) and the risks predicted by the model, we use the *Odds Ratio* (OR). The OR is a statistical measure that quantifies the strength and direction of the association between two events. An OR greater than 1 suggests a positive association, while an OR less than 1 indicates a negative association. An OR of 1 means there is no association between the two events. In this context, the OR quantifies how a specific factor influences the odds of a particular a risk outcome. It expresses how much greater the odds are for one category of the factor compared to another.

Example 1 In Table 5, we have recorded some *a posteriori* probabilities for the patient in Example 1, whose characteristics are detailed in Table 4. These probabilities are provided for various pairs of α and β values. We assume that the patient has a therapeutic limit, although the specific limit is unknown. As observed, increasing α with a fixed β, or decreasing β with a fixed α, results in a decrease in the *a posteriori* probability of discharge, while the probabilities of exitus and icu increase. The probability of icu is the highest, and the confidence level (highest probability) is highlighted in bold.

How does the specific type of therapeutic limit affect the *a posteriori* probabilities we have obtained? Table 6 shows the results for the two possible therapeutic limits, highlighting the differences: when the therapeutic limit is NRB (non-rebreather mask), the probability assigned to icu remains very low across the entire range of parameter values. In contrast, when the therapeutic limit is NIMV (non-invasive mechanical ventilation), this probability increases up to 20% (when $\alpha = \beta = 20$).



(a) TC metric. NB model

(b) TC metric. NB model

(c) BA metric. SVM model

(d) BA metric. NB model

Figure 5. Heatmaps depicting the mean TC and BA values as functions of $\alpha = 2, \ldots, 20$ and $\beta = 2, \ldots, \alpha$ for the best models, with the MTh meta-algorithm. The median values are shown in white, with lower values in blue and higher values ir red.

Demographic	Age: 50-65
Vital signs	Charlson Index: 1
	O2 saturation: low
Symptoms	asthenia: yes
	dyspnoea: yes
Blood test	c-reactive protein: 100-150
	d-dimer: 300-500
	lactate: 250-350
Previous treat.	immunosuppressants: yes
	statins: yes

Table 4. Patient characteristics for Example 1: Case with a therapeutic limit.

Probabilities Example 1			Therapeutic limit Limit type: unknown discharge	exitus	icu
a priori ⟶			**60.58%**	**36.76%**	**2.66%**
	α	β	discharge	exitus	icu
a posteriori	2	2	**98.78%**	0.48%	0.74%
	10	2	**97.12%**	1.21%	1.67%
	10	5	**96.25%**	1.51%	2.24%
	10	10	**94.90%**	1.93%	3.17%
	20	2	**95.36%**	2.09%	2.55%
	20	10	**93.51%**	2.65%	3.84%
	20	20	**91.50%**	3.11%	5.39%

Table 5. *A priori* and *a posteriori* probabilities for the patient in Example 1. Limit type unknown.

Probabilities Example 1			Therapeutic limit Limit type: NIMV			Limit type: NRB		
	α	β	discharge	exitus	icu	discharge	exitus	icu
a posteriori	2	2	**96.61%**	0.47%	2.92%	**99.50%**	0.37%	0.12%
	10	2	**94.99%**	0.75%	4.26%	**98.11%**	1.46%	0.42%
	10	5	**91.51%**	1.12%	7.37%	**97.97%**	1.57%	0.46%
	10	10	**86.40%**	1.50%	12.10%	**97.73%**	1.74%	0.53%
	20	2	**93.28%**	1.08%	5.64%	**96.52%**	2.79%	0.70%
	20	10	**85.65%**	1.72%	12.63%	**96.22%**	2.99%	0.79%
	20	20	**78.01%**	1.99%	20.00%	**95.88%**	3.22%	0.90%

Table 6. *A posteriori* probabilities for the patient in Example 1 based on the known type of therapeutic limit.

Table 6 reveals a striking contrast in the estimated *a posteriori* risks of exitus and icu depending on the type of therapeutic limit. When the limit is NRB, the probabilities of exitus closely resemble those when the limit type is unknown, with the probability of discharge increasing at the expense of the icu probability, which remains notably low. Interestingly, in this scenario, the icu probability is even lower than that of exitus, which differs from the situation when the limit type is unknown. Conversely, with NIMV, the estimated *a posteriori* risk of exitus is low, while the risk of icu significantly increases at the expense of the discharge probability. This underscores the substantial impact of the therapeutic limit type on the model's risk predictions.

The heatmaps in Appendix C display the *a posteriori* risks of exitus and icu for this example. These heatmaps not only provide the magnitude of these probabilities through colour scales, which may vary from one heatmap to another, but also reveal discernible trends. They illustrate how the risks change as we increase the values of α, and for each α, as we increase β. Importantly, this trend analysis depends on the specific type of therapeutic limit. Examining these heatmaps allows us to gain insights into how risk probabilities are affected by different parameter combinations and how these variations rely on the particular scenario at hand.

By focusing on regions with higher (red) or lower (blue) outcome values, we can closely examine the heatmaps and draw inferences about which combinations of parameters lead to better or worse outcomes. This allows us to perform a **sensitivity analysis** using the heatmaps as a valuable tool to understand how changes in parameters impact risk assessments for COVID-19 patients. For instance, in Figure 10, we observe that the color bands corresponding to the NRB therapeutic limit are nearly vertical for the risk of death. This suggests that the risk is sensitive to the value of α, increasing as α increases, while remaining relatively stable concerning β. In contrast, the nearly horizontal color bands, such as those for the NIMV therapeutic limit in ICU admission risk, indicate sensitivity to the value of β, with the risk increasing as β increases but showing robustness with respect to α. These observations align with the cost matrix expression in (1), providing valuable insights into how the predicted risks behave based on the specific therapeutic limit type.

Next, we calculate the estimated Odds Ratios (OR) based on the probabilities assigned by the model for the patient in Example 1. (We can estimate the OR values from the provided probabilities, but without sample sizes or additional information, meaningful confidence intervals for the OR cannot be calculated). We consider the type of therapeutic limit as factor that could be associated with the risk of death or ICU admission. Table 7 presents the computed OR values for icu and exitus for a patient corresponding to Example 1. These OR values are calculated when the therapeutic limit is NIMV compared to NRB. For example, with $\alpha = \beta = 2$, the OR in favor of icu is calculated as follows: $\dfrac{0.0292/(1-0.0292)}{0.0012/(1-0.0012)} = 25.03516 \approx 25.035$.

Patient Example 1	$\alpha = 2$	$\alpha = 10$			$\alpha = 20$		
OR NIMV w.r.t. NRB	$\beta = 2$	$\beta = 2$	$\beta = 5$	$\beta = 10$	$\beta = 2$	$\beta = 10$	$\beta = 20$
icu	25.035	10.550	17.217	25.835	8.479	18.154	27.528
exitus	1.272	0.510	0.710	0.860	0.380	0.568	0.610

Table 7. OR for ICU admission risk (1st row) and death risk (2nd row) for the patient in Example 1 with therapeutic limit NIMV, compared to NRB, across different α and β values, based on probabilities from Table 6.

		Therapeutic limit								
		Limit type: unknown			Limit type: NIMV			Limit type: NRB		
α	β	disch.	exitus	icu	disch.	exitus	icu	disch.	exitus	icu
2	2	**83.88%**	15.45%	0.67%	**90.88%**	7.74%	1.38%	**79.59%**	20.23%	0.18%
10	2	**52.57%**	46.38%	1.05%	**71.02%**	26.16%	2.88%	44.17%	**55.58%**	0.25%
10	5	**52.75%**	46.18%	1.07%	**70.86%**	26.16%	2.98%	44.24%	**55.51%**	0.25%
10	10	**53.03%**	45.86%	1.11%	**70.60%**	26.15%	3.25%	44.35%	**55.40%**	0.25%
20	2	35.97%	**63.13%**	0.90%	**56.35%**	40.70%	2.95%	28.40%	**71.40%**	0.20%
20	10	36.90%	**62.18%**	0.94%	**57.65%**	39.09%	3.26%	28.65%	**71.15%**	0.20%
20	20	37.98%	**61.03%**	0.99%	**58.88%**	37.48%	3.64%	28.96%	**70.84%**	0.20%

Table 8. *A posteriori* probabilities for the patient in Example 2 based on the known type of therapeutic limit.

In Table 7, when OR > 1, it indicates that the NIMV therapeutic limit is a risk factor compared to NRB. This is true for ICU admission risk across all values of α and β. However, it only applies to the risk of death when $\alpha = \beta = 2$. For the other tested parameter values, NIMV serves as a protective factor against NRB for the risk of death.

Clinically, this can be explained as follows. When a patient is placed on NRB, it indicates they can receive oxygen support but are not on mechanical ventilation. This implies that the patient's initial condition is more fragile, and a proactive decision is made to avoid escalating treatment, including ICU admission and ventilation techniques. Such patients are at a higher risk if subjected to intensive care and ventilation maneuvers, which could potentially increase their mortality rate. Consequently, they have a lower likelihood of ICU admission, which, unfortunately, also entails higher mortality rates. On the other hand, NIMV implies that if the patient's respiratory distress has reached a point where mechanical support is necessary, they are typically provided with non-invasive ventilation within the ICU, with an emphasis on avoiding more invasive treatments. Patients with no therapeutic limit may undergo invasive mechanical ventilation (intubation in the ICU) if necessary. Therefore, the concept of a NIMB therapeutic limit reflects a deliberate approach to critical care, balancing life-saving interventions and minimizing invasiveness.

The results in Table 7 align with clinical intuition and offer valuable quantification of these trends. This quantification is essential and serves various purposes. While healthcare professionals may have a general understanding of how different factors affect patient outcomes, quantifying these effects provides an objective, evidence-based assessment. It enables healthcare providers to make more informed decisions about the level of care and interventions required for specific patients. Resource allocation is particularly crucial in healthcare settings, especially during a pandemic. Understanding how various factors impact patient outcomes aids in the efficient distribution of resources. In summary, quantifying the influence of therapeutic factors such as the type of therapeutic limit (NRB and NIMV) on patient outcomes, as demonstrated in Table 7, is fundamental for evidence-based medicine. It enhances the precision of clinical decision-making, resource allocation, risk communication, healthcare research, and overall healthcare system management, ultimately resulting in improved patient care.

Example 2 Now, let us consider a patient with the same attributes as the one in Example 1 but with a Charlson Index of $4 - 5$. Clinically, a higher Charlson score indicates an increased risk of adverse outcomes. The resulting *a posteriori* probabilities for this patient are detailed in Table 8.

Using the values in Table 8, we can calculate Odds Ratios (OR) similar to our approach in Example 1. It is particularly insightful to assess the "Charlson Index effect". This involves understanding how a higher Charlson Index (as in Example 2) versus a lower one (as in Example 1) influences the patient's risks, keeping all other attributes constant. Evaluating the impact of the Charlson Index on death and ICU admission risk assessments helps healthcare professionals make informed decisions about treatment and care strategies. It also assists in the efficient allocation of resources, especially in the context of COVID-19 patient care.

This outcomes are consistent with those in Table 6 and align with our qualitative expectations. However, the true value of the model lies in its ability to quantify these expectations. Notably, for high α values, the model assigns the exitus category to the patient in Example 2, even when the therapeutic limit is of type NRB (or when

Demographic	Age: 50-65
Vital signs	O2 saturation: low
Symptoms	dyspnoea: yes
Blood test	c-reactive protein: 100-150
	d-dimer: 300-500
	lactate: 250-350
Previous treat.	statins: yes

Table 9. Patient characteristics for Example 3.

Probabilities Example 3		discharge	exitus	icu
No therapeutic limit		48.79%	1.45%	**49.76%**
Therapeutic limit	unknown	**70.50%**	21.62%	7.88%
	NIMV	**64.07%**	10.03%	25.80%
	NRB	**67.67%**	30.75%	1.58%

Table 10. *A posteriori* probabilities for the patient in Example 3 with $\alpha = \beta = 20$.

the limit type is unknown). In contrast, the patient in Example 1 was consistently assigned to the discharge category under the same conditions. This demonstrates the model's ability to capture subtle variations in risk assessment based on differing patient characteristics. For further insights, please refer to the corresponding heatmaps in Appendix C.

Example 3 To evaluate the impact of the presence and type of a therapeutic limit, we compare the estimated risks for a patient with the characteristics outlined in Table 9. This comparison hinges on whether the patient has a therapeutic limit, its type if applicable, and whether it is known. We have selected $\alpha = \beta = 20$ for this analysis. The results are summarized in Table 10.

Given the vast number of potential scenarios, we will not explore further examples in this section. However, consider cases where multiple features influence outcomes simultaneously, such as age and the Charlson Index. In these situations, one feature might be a risk factor (e.g., a high Charlson Index) while another could act as a protective factor (e.g., low age). Evaluating the combined effect of a risk factor and a protective factor is complex without a suitable quantitative model. Our model provides a valuable tool for assessing how certain characteristics (protective factors) can mitigate the impact of risk factors on the likelihood of ICU admission or death.

Discussion and Conclusions

This study presents a comprehensive evaluation of the impact of therapeutic limits on predictive models for patient mortality and ICU admission risks, particularly within the context of COVID-19. By developing predictive models tailored to patient subgroups with and without therapeutic limits, we achieved significant improvements in risk prediction accuracy through the introduction of the novel MTh meta-algorithm. This innovative algorithm extends thresholding to multiclass settings and offers a cost-sensitive approach to predictive modelling, marking a significant progression in the field. The key findings and implications of the study are:

1. **Impact of therapeutic limits:** The choice and type of therapeutic limits, such as Non-Rebreather Mask (NRB) or Non-Invasive Mechanical Ventilation (NIMV), play a critical role in patient outcomes. This choice reflects the initial assessment of patient frailty, with NIMV typically indicating a less fragile condition, allowing for more intensive interventions if necessary. This often results in a higher risk of ICU admission but also an improved chance of survival. On the other hand, NRB is generally associated with a more fragile condition, preventing ICU admission and the use of invasive measures. While this may suffice for less severe cases, it can lead to higher mortality rates for patients in more critical conditions.

   This complex interplay of clinical factors underscores the challenges in clinical decision-making, where therapeutic interventions must carefully balance the severity of the patient's condition, the available resources, and the expected outcomes. The ML models developed for each patient subgroup demonstrate how different therapeutic limits correlate with distinct probabilities of ICU admission and mortality, thereby supporting clinicians in making more informed decisions and optimizing resource allocation.

2. **Explanatory and predictive insights:** The integration of Bayesian Networks as explanatory models provides a clearer understanding of the interdependencies among patient variables, thereby supporting more nuanced and informed clinical decision-making. Our predictive model effectively quantifies the impact of patient

characteristics on risk predictions. The combination of predictive and explanatory modelling offers a robust framework for addressing the complexities of patient care.

3. **Introduction of MTh meta-algorithm:** MTh extends the concept of thresholding from binary to multi-class settings by adopting a cost-sensitive approach, focusing on the Total Cost (TC) metric as a behavioral indicator. This novel approach is particularly effective in handling multiclass classification problems and imbalanced datasets. Additionally, the adaptability of the MTh algorithm by updating the cost matrix to reflect changing circumstances, makes it a highly responsive and valuable tool in the evolving landscape of healthcare.Collectively, these findings highlight the importance of integrating advanced ML techniques into healthcare to improve patient outcomes and optimize clinical decision-making. However, several potential biases and limitations should be considered: (1) the model's performance is based on a specific dataset with a relatively small number of cases, which may limit its generalizability to broader patient populations; (2) the model's sensitivity to the chosen parameters (α and β), representing misclassification costs, may affect the stability of risk predictions, and imprecise specifications of these parameters could undermine the reliability of the results; and (3) inaccuracies in identifying therapeutic limits could compromise the integrity of risk assessments.

These limitations underscore the need for ongoing research and refinement to enhance the model's robustness and applicability across different healthcare scenarios. Future research should focus on extending the model's applicability by incorporating more diverse patient populations and comorbidities, thereby enhancing its generalizability, robustness, and relevance. Additionally, further exploration of the effects of parameters α and β on model reliability is essential. To address potential errors in the assignment of appropriate therapeutic limits, we plan to implement targeted training for medical personnel. Lastly, integrating the predictive model into electronic health records (EHR) systems could facilitate real-time risk assessments, improving clinical workflows and decision-making efficiency. This integration would represent a significant step forward in bringing predictive modelling into everyday clinical practice.

In summary, this work not only offers a holistic framework that integrates predictive and explanatory modelling to deliver actionable insights into patient outcomes, but also marks the introduction of the MTh meta-algorithm, a novel and significant advancement in predictive modelling. By advancing the application of ML in healthcare –particularly through the development and validation of the MTh meta-algorithm– we contribute to the ongoing evolution of personalized medicine. Our approach is cost-sensitive and employs rigorous model evaluation techniques. Its successful implementation demonstrates potential to enhance clinical decision-making, offering a promising path toward more tailored and effective patient care in the future.

Data availability

The data supporting these findings are restricted due to ethical and legal constraints. Access to these data may be granted upon request to the Ethics Committee of Bellvitge University Hospital (Barcelona, Spain), which will evaluate each request individually. To inquire about data access, please contact the Research Support Unit at Bellvitge University Hospital through the following email address: clinicalresearchwindow@bellvitgehospital.cat, or visit the website at www.bellvitgehospital.cat/clinicalresearch

Appendix A: Overview of patient characteristics in the dataset

Variables highlighted in italic are exclusively utilized in constructing the predictive model for the data subset of patients with a therapeutic limit, while those highlighted in bold italic are specific to the data subset of patients without a therapeutic limit. Variables in black are used for both data subsets. Percentages refer to the entire dataset.

Demographic

age	< 50 17.37%, 50–65 25.97%, 65–75 22.34%, 75–85 23.08%, ≥ 85 11.24%
sex	man 58.42%, woman 41.58%

Initial assessment

limit type	NRB 68.26%, NIMV 31.74%

Therapeutic limit can be either NRB (non-rebreather mask) or NIMV (non-invasive mechanical ventilation).

Vital signs

Charlson Index	0 13.00%, 1 13.41%, 2–3 27.48%, 4–5 22.96%, > 5 21.39%, unknown 1.76%
consciousness	normal 80.75%, abnormal 5.15%, unknown 14.10%
O_2 saturation	normal 51.43%, low 26.53%, hypoxia 14.87%, unknown 7.17%

The Charlson Index is a medical score designed to predict 10-year survival in patients with multiple comorbidities. It ranges from 0 to 29.

Symptoms

abdominal pain	yes 4.61%, no 75.67%, unknown 19.72%
anosmia & ageusia	both 6.31%, ageusia 3.36%, anosmia 2.26%, neither 84.29%, unknown 3.78%
arthromyalgia	yes 28.26%, no 69.18%, unknown 2.56%
asthenia	yes 24.99%, no 59.10%, unknown 15.91%
cephalea	yes 12.58%, no 85.10%, unknown 2.32%
confusional syndrome	yes 5.21%, no 77.90%, unknown 16.89%
dyspnoea	yes 46.46%, no 36.73%, unknown 16.81%
nauseas	yes 9.40%, no 71.15%, unknown 19.45%
rhinorrhea	yes 4.52%, no 88.55%, unknown 6.93%
thoracic pain	yes 6.37%, no 73.85%, unknown 19.78%

Rhinorrhea refers to nasal congestion, *anosmia* signifies a loss of smell, and *ageusia* loss of taste. *Arthromyalgia* indicates muscle or joint pain. *Dyspnoea* represents shortness of breath, *asthenia* denotes fatigue, and *cephalea* corresponds to headache.

Blood test

c-reactive protein	< 20 11.01%, 20–50 18.62%, 50–100 22.31%, 100–150 15.97%, $\geq$ 150 23.91%, unknown 8.18%
d-dimer	$<$ 300 15.14%, 300–500 16.18%, 500–1000 20.14%, $\geq$ 1000 22.46%, unknown 26.08%
lactate	$<$ 250 12.49%, 250–350 19.54%, 350–450 10.56%, $\geq$ 450 11.63%, unknown 45.78%
lymphocytes	$<$ 0.75 29.54%, 0.75–1 22.07%, 1–1.5 26.65%, $\geq$ 1.5 15.76%, unknown 5.98%

d-dimer is a byproduct of blood clots, while *C-reactive protein* is a protein that elevates in response to inflammation.

Comorbidities

coronary artery disease	yes 5.92%, no 77.75%, unknown 16.33%
dementia	yes 7.29%, no 92.38%, unknown 0.33%
diabetes	yes-complications 5.35%, yes-no-complications 19.01%, no 75.43%, unknown 0.21%
hemiplegia	yes 0.98%, no 98.75%, unknown 0.27%
ictus	yes 6.84%, no 92.83%, unknown 0.33%
mild kidney failure	yes 9.28%, no 71.68%, unknown 19.04%

Hemiplegia refers to the paralysis of one half of the body in a patient.

Previous treatments

acetylsalicylic acid	yes 11.15% , no 59.43%, unknown 29.42%
antibiotic	yes 11.54%, no 84.12%, unknown 4.34%
anticoagulants	yes 9.46%, no 88.85%, unknown 1.69%
biological therapies	yes 0.51%, no 85.99%, unknown 13.50%
hydroxychloroquine	yes 0.72%, no 93.81%, unknown 5.47%
immunosuppressants	yes 2.56%, no 84.03%, unknown 13.41%
statins	yes 25.52%, no 58.30%, unknown 16.18%

Appendix B: Arc strength and influential features

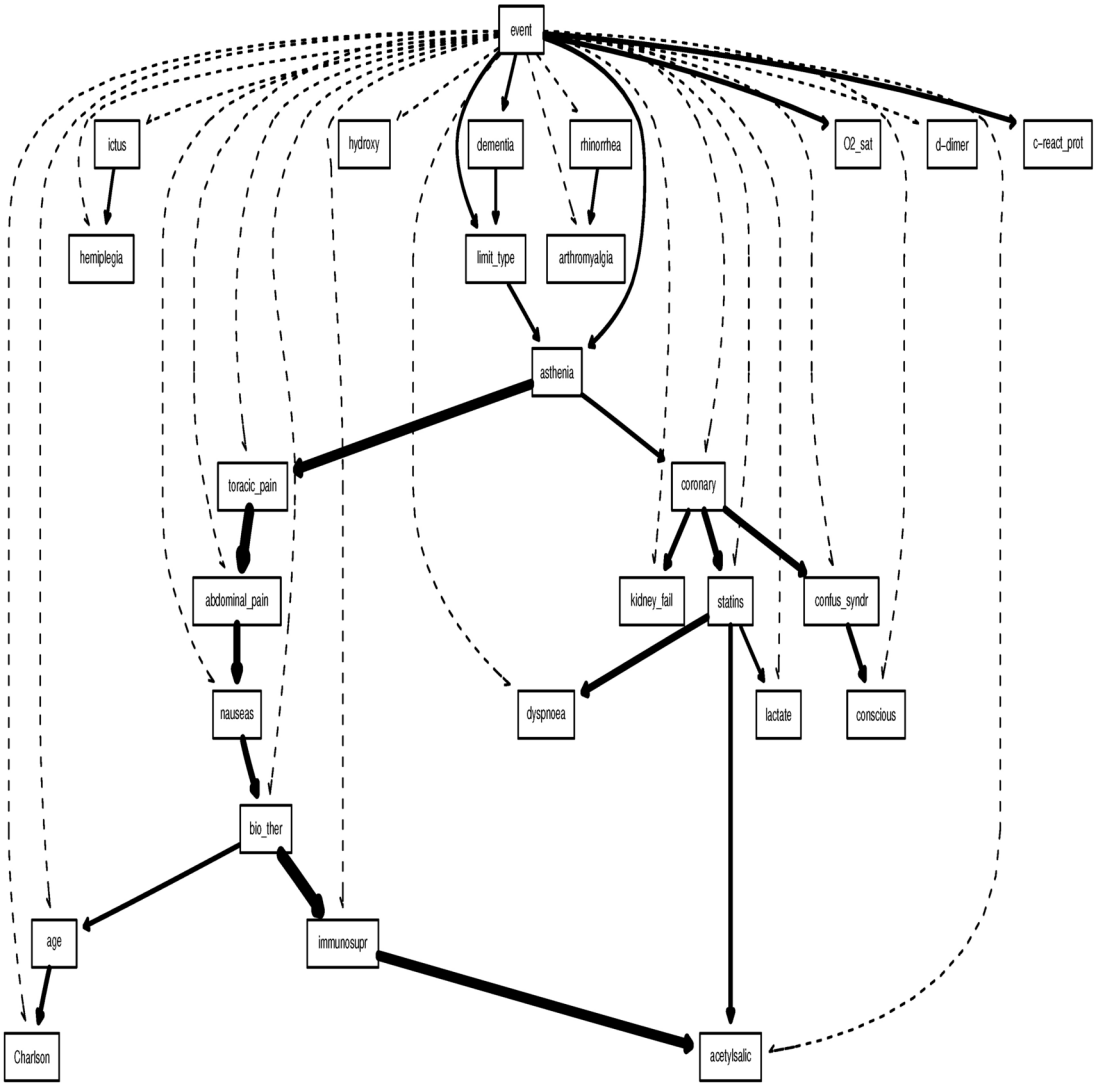


Figure 6. Arcs strength in the DAG of the Augmented Naive model for the "therapeutic limit" data subset.

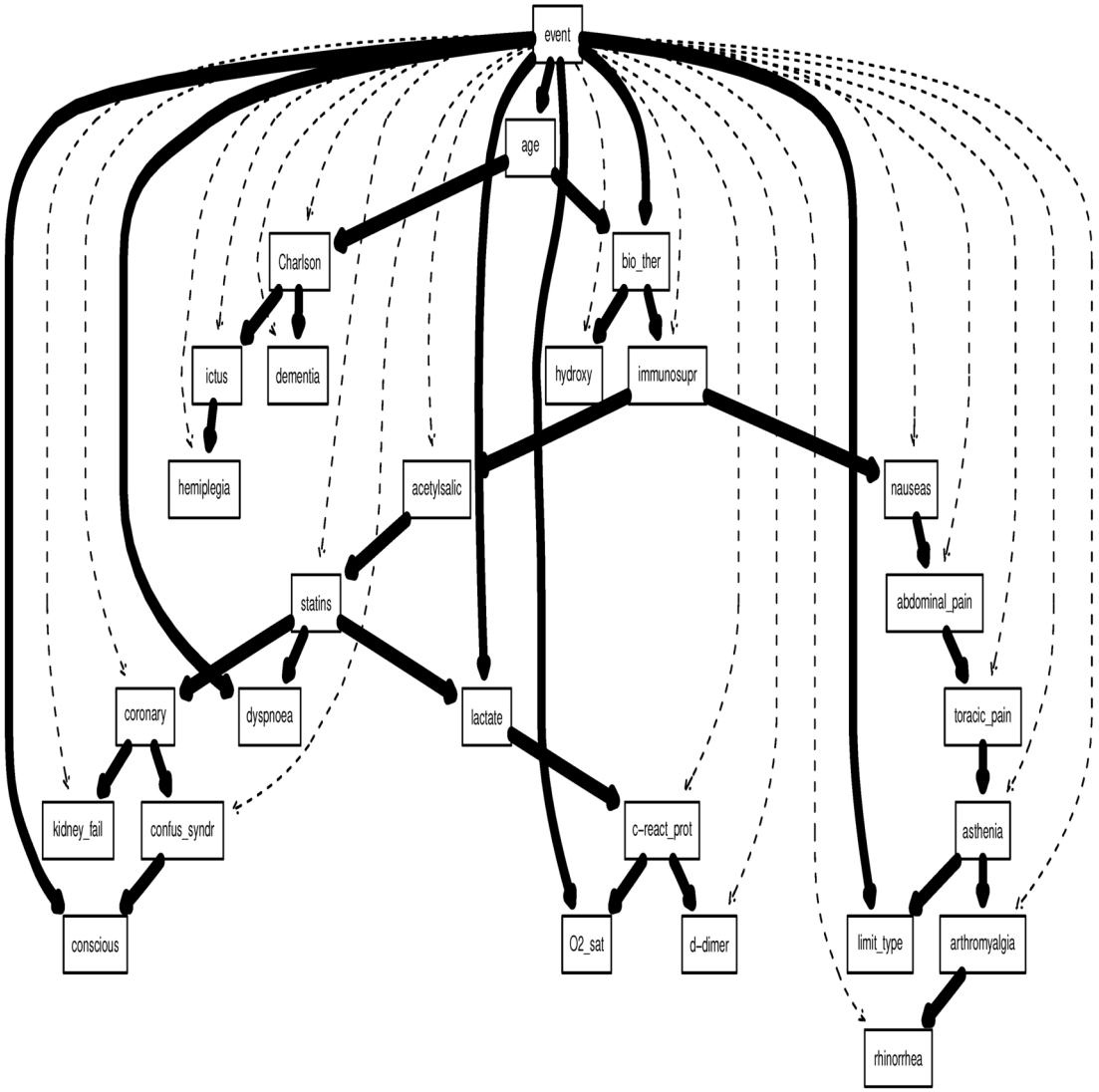


Figure 7. Arcs strength in the DAG of the TAN model for the "therapeutic limit" data subset.

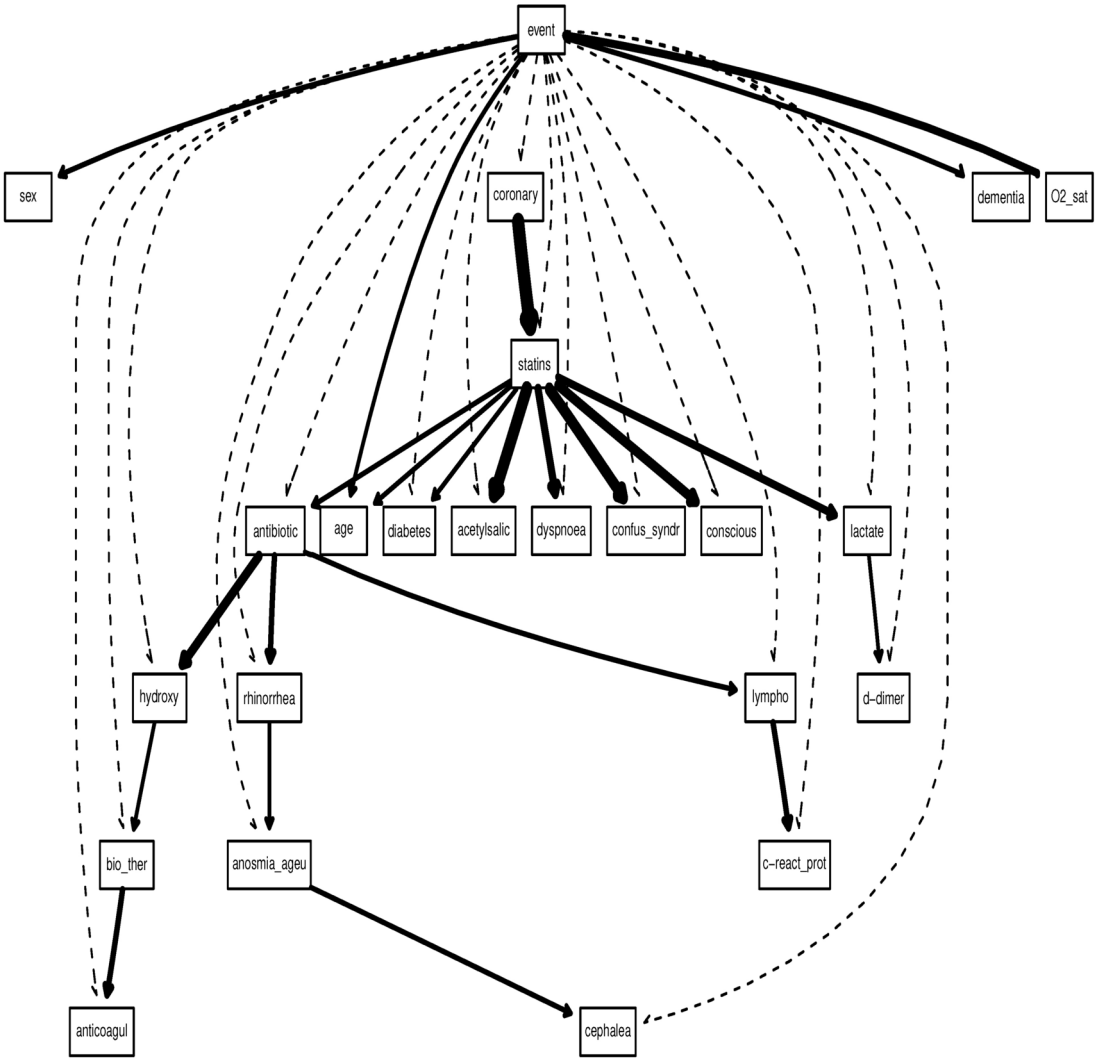


Figure 8. Arcs strength in the DAG of the Augmented Naive model for the "no-therapeutic limit" data subset.

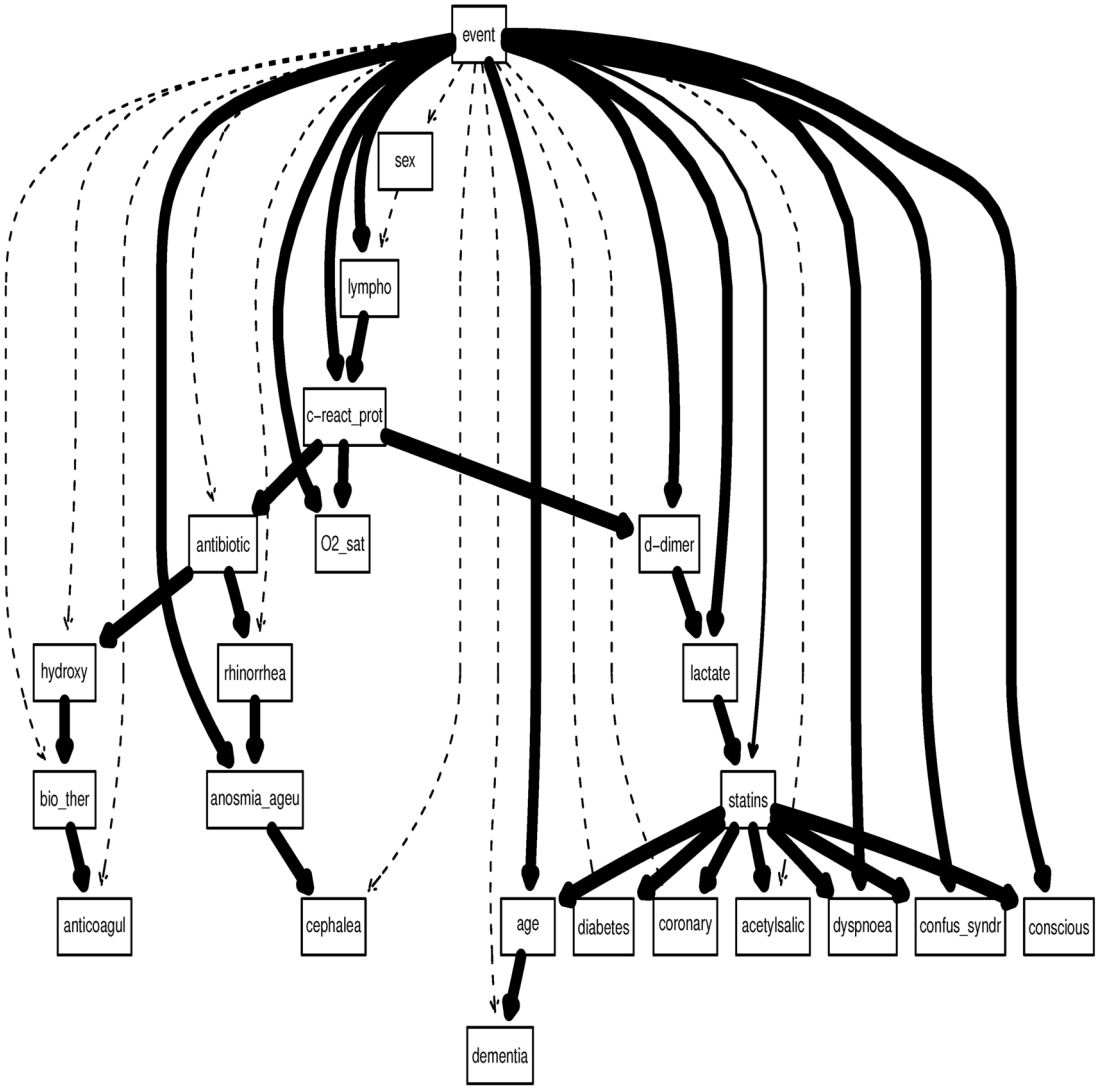


Figure 9. Arcs strength in the DAG of the TAN model for the "no-therapeutic limit" data subset.

Therapeutic limit		NB	AN	TAN
Demographic	age	1.0262×10^{-26}	6.8210×10^{-19}	1.0262×10^{-26}
Initial assessment	limit type	2.1028×10^{-14}	5.1537×10^{-10}	1.0367×10^{-32}
Vital signs	Charlson Index	1.4510×10^{-23}		
	consciousness	1.8351×10^{-16}	4.2450×10^{-9}	
	O2 saturation	4.7308×10^{-21}		8.0103×10^{-9}
Symptoms	abdominal pain	9.3183×10^{-10}		
	arthromyalgia	$1.1429 \times 10^{-2*}$		
	asthenia	1.3104×10^{-11}	5.4121×10^{-30}	4.95774×10^{-4}
	confusional syndrome	1.3891×10^{-7}	1.8365×10^{-5}	
	dyspnoea	9.2462×10^{-9}	2.4929×10^{-7}	
	nauseas	6.2838×10^{-11}		
	rhinorrhea	$1.9114 \times 10^{-3**}$		$3.4834 \times 10^{-2*}$
	thoracic pain	9.6998×10^{-11}	$2.2364 \times 10^{-3**}$	
Blood test	c-reactive protein	9.1693×10^{-12}		6.0460×10^{-4}
	d-dimer	1.2327×10^{-6}		$2.3893 \times 10^{-3**}$
	lactate	2.3629×10^{-18}	5.3051×10^{-17}	
Comorbidities	dementia	8.5090×10^{-9}		
	hemiplegia	9.2061×10^{-4}		
	mild kidney failure	$3.9761 \times 10^{-2*}$	$4.7419 \times 10^{-2*}$	
Previous treat.	acetylsalicylic acid	5.3530×10^{-12}		
	biological therapies	8.0247×10^{-14}	8.3223×10^{-4}	3.1345×10^{-7}
	hydroxychloroquine	$2.6519 \times 10^{-3**}$		$7.4428 \times 10^{-3**}$
	immunosuppressants	3.0180×10^{-14}		
	statins			$7.7630 \times 10^{-3**}$

Table 11. Therapeutic limit. Most influential features for risk prediction, with p-values indicating strength of influence (lower p-values indicate stronger influence). Unless otherwise indicated by a superscript * (5%) or ** (1%), the statistical significance of p-values is $1‰$.

No-therapeutic limit		NB	AN	TAN
Demographic	age	2.8061×10^{-51}	4.1440×10^{-41}	
	sex	1.1142×10^{-5}		
Vital signs	consciousness	1.8257×10^{-11}	9.5856×10^{-13}	
	O2 saturation	3.1983×10^{-46}		2.9497×10^{-24}
Symptoms	anosmia & ageusia	1.3770×10^{-11}	5.3000×10^{-8}	
	cephalea	2.2856×10^{-6}		
	confusional syndrome	2.4702×10^{-7}	1.8388×10^{-7}	
	dyspnoea	4.9554×10^{-13}	3.7987×10^{-10}	
	rhinorrhea	1.1796×10^{-2} *	4.1275×10^{-3} **	
Blood test	c-reactive protein	2.8987×10^{-31}	8.6838×10^{-12}	
	d-dimer	2.0911×10^{-16}	1.8923×10^{-6}	1.0311×10^{-7}
	lactate	1.4145×10^{-23}	1.5090×10^{-22}	1.5801×10^{-11}
	lymphocytes	6.3165×10^{-21}	6.5585×10^{-16}	8.4156×10^{-16}
Comorbidities	dementia	3.1484×10^{-15}		
	diabetes	1.7290×10^{-4}	1.3544×10^{-3} **	
Previous treat.	acetylsalicylic acid	1.7857×10^{-3} **	7.4592×10^{-3} **	
	antibiotic	2.65134×10^{-2} *	3.0119×10^{-2} *	
	anticoagulants	7.3877×10^{-6}	4.2265×10^{-4}	
	statins	7.0450×10^{-5}	2.2817×10^{-3} **	6.7093×10^{-6}

Table 12. No-therapeutic limit. Most influential features for risk prediction, with p-values indicating strength of influence (lower p-values indicate stronger influence). Unless otherwise indicated by a superscript * (5%) or ** (1%), the statistical significance of p-values is $1‰$.

Appendix C: Heatmaps for risk predictions

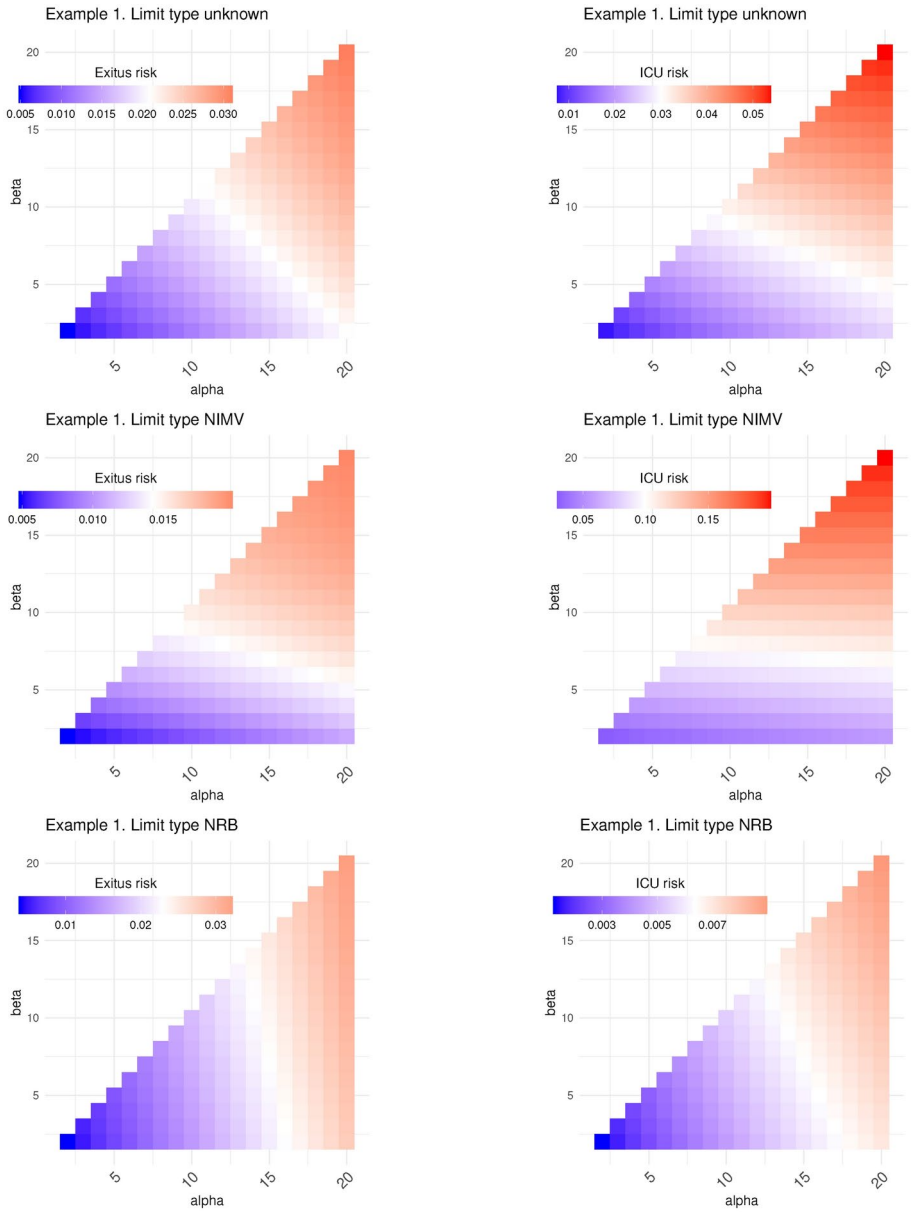


Figure 10. Heatmaps showing the *a posteriori* probabilities for the patient in Example 1: probability of exitus (left) and icu (right) as functions of $\alpha = 2, \ldots, 20$ and $\beta = 2, \ldots, \alpha$, differentiated by therapeutic limit type. White represents the median value, with low probabilities in blue and high probabilities in red.

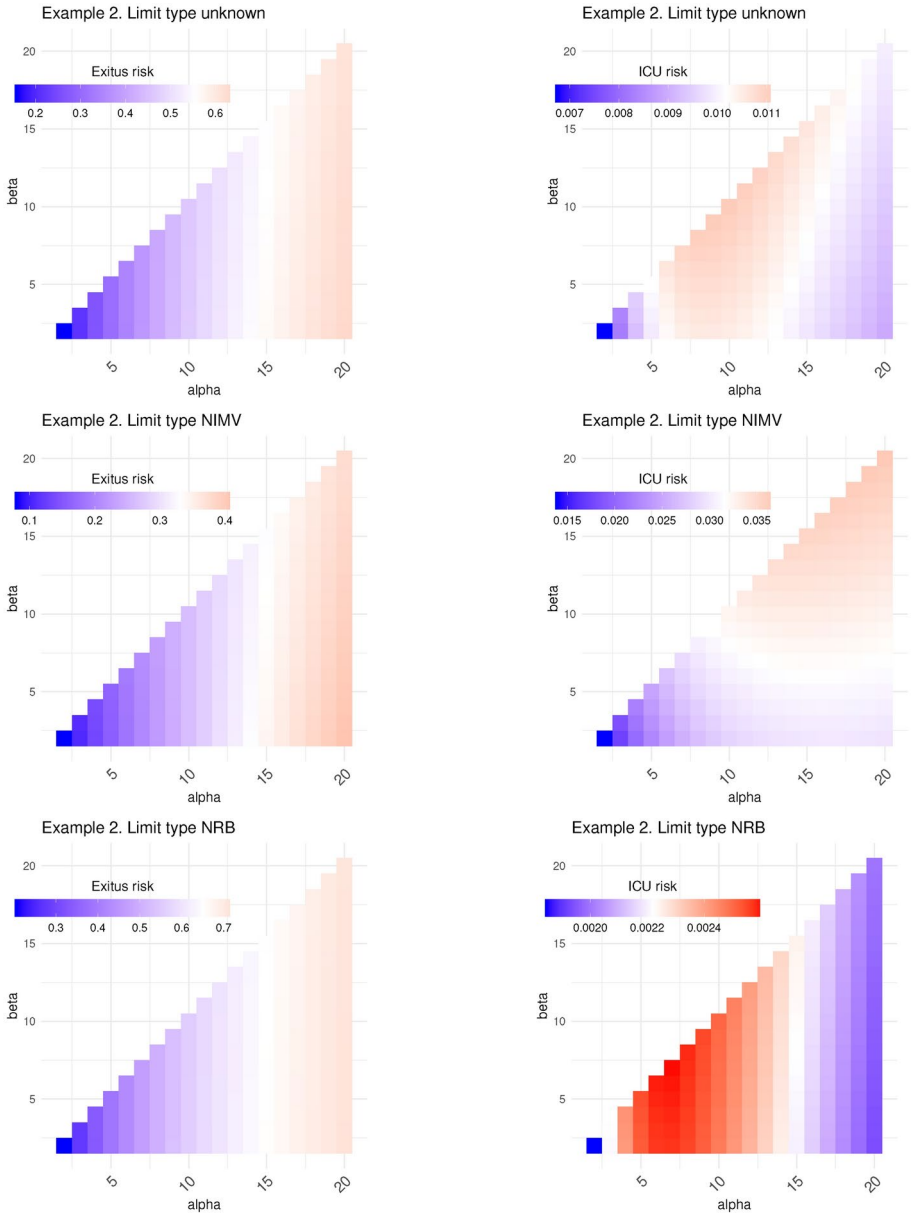


Figure 11. Heatmaps showing the *a posteriori* probabilities for the patient in Example 2: probability of exitus (left) and icu (right) as functions of $\alpha = 2, \ldots, 20$ and $\beta = 2, \ldots, \alpha$, differentiated by therapeutic limit type. White represents the median value, with low probabilities in blue and high probabilities in red.

Appendix D: The Multi-Thresholding meta-algorithm, MTh

In this appendix, we present and establish some properties of Algorithm 1.

Remark 1

Bayes Minimum Risk (BMR) is a concept used in classification problems, particularly in the context of Bayesian Decision Theory and decision-making under uncertainty. BMR is especially valuable in situations where classification errors have different costs or consequences. This approach involves making predictions by selecting the class that minimizes the risk (or expected cost) associated with incorrectly predicting that class. Formally and with our notations:

$$c^*_{\mathrm{BMR}} = c_h \quad \text{with} \quad h = \arg\min_{i=1,\ldots,r} \omega_i \,.$$

In contrast, the MTh meta-algorithm is based on selecting the class that maximizes the adjusted probabilities obtained by dividing the original probabilities by the corresponding risks. When the original probabilities are equal ($p_1 = \cdots = p_r$), both criteria coincide, meaning that $c^* = c^*_{\mathrm{BMR}}$.

The following result demonstrates that, in the binary case, MTh essentially acts as a "thresholding" method, modifying the conventional threshold of 0.5 in cost-insensitive approaches. This modified threshold determines which class label is assigned.

Proposition 1

In the binary case where $r = 2$, MTh assigns class label c_1 when $p_1 > \tau$, and it assigns class label c_2 when $p_2 > 1 - \tau$ (otherwise, there is no clear evidence for either class, and a tie-breaking mechanism should be implemented). The value of this "threshold", denoted as τ, is defined as follows with $\mu = m_{21}/m_{12}$:

$$\tau = \begin{cases} 1/2 & \text{if } m_{21} = m_{12} \\ 0 & \text{if } m_{21} > m_{12} = 0 \\ 1 & \text{if } m_{12} > m_{21} = 0 \\ \dfrac{-1+\sqrt{\mu}}{\mu-1} & \text{otherwise.} \end{cases}$$

Proof

MTh assigns class label c_1 if $\widetilde{p}_1 > \widetilde{p}_2$. Since $\omega_1 = m_{12}\, p_2$ and $\omega_2 = m_{21}\, p_1$, we can express this condition as follows:

$$\widetilde{p}_1 > \widetilde{p}_2 \Leftrightarrow \frac{p_1}{m_{12}\, p_2} > \frac{p_2}{m_{21}\, p_1} \Leftrightarrow (m_{21} - m_{12})\, p_1^2 + 2\, m_{12}\, p_1 - m_{12} > 0\,. \tag{D1}$$

Here, we have used the fact that $p_2 = 1 - p_1$.

Notably, when $m_{12} = m_{21}$, the condition (D1) simplifies to $\Leftrightarrow p_1 > 1/2$, aligning with the cost-insensitive scenario as expected.
Let's assume, for now, that $m_{21} > m_{12}$. If $m_{12} = 0$, equation (D1) is equivalent to $p_1 > 0$. In this case, when there is no cost associated with misclassifying and instance of class c_2 as c_1 (but there is a cost for the reverse misclassification), MTh always assigns class label c_1, which is a logical outcome. Otherwise, if we introduce the ratio $\mu = m_{21}/m_{12}$, which is greater than 1, equation (D1) becomes equivalent to

$$(\mu - 1)\, p_1^2 + 2\, p_1 - 1 > 0\,. \tag{D2}$$

The roots of this quadratic equation are:

$$\frac{-1-\sqrt{\mu}}{\mu-1} < 0 < \tau = \frac{-1+\sqrt{\mu}}{\mu-1} < \frac{1}{2}\,.$$

Therefore, equation (D2) holds if and only if $p_1 > \tau$.

The case $m_{21} < m_{12}$ is analogous to the previous case. In particular, if $m_{21} = 0$, equation (D1) is equivalent to $(p_1 - 1)^2 < 0$, which is not true for any value of p_1, leading to $\tau = 1$. The interpretation here is that if there is no cost associated with misclassifying an instance of class c_1 as c_2 (but there is a cost for the reverse misclassification), MTh logically never assigns class label c_1. Now, in the case where $0 < m_{21} < m_{12}$, which implies that $\mu = m_{21}/m_{12} < 1$, the two roots of equation (D2) are:

$$\frac{1}{2} < \tau = \frac{-1+\sqrt{\mu}}{\mu-1} < 1 < \frac{1+\sqrt{\mu}}{1+\mu}\,.$$

Consequently, equation (D2) (or equivalently, (D1)) holds if and only if $p_1 > \tau$ due to the fact that $\mu < 1$. □

Corollary 1

In the binary case where $r = 2$, MTh assigns class label c_1 if $p_1 > \tau$, where the "threshold" τ satisfies the condition:

$$\begin{cases} 0 < \tau < 1/2 & \text{if } \; m_{21} > m_{12} > 0 \\ 1/2 < \tau < 1 & \text{if } \; 0 < m_{21} < m_{12}\,. \end{cases}$$

That is, if the cost associated with misclassifying an instance of class c_i is greater than the cost of the opposite error, MTh assigns class label c_i with a higher probability than the other class label.

Corollary 2

In the binary case where $r = 2$, the "threshold" τ obtained with the MTh algorithm is a continuous and decreasing function of μ within the interval $[0, +\infty)$. As μ approaches infinity, τ tends towards zero.

Proof

The "threshold" τ obtained with the MTh algorithm where $r = 2$ is given by the following function of μ in $[0, +\infty)$:

$$\tau = f(\mu) = \begin{cases} \dfrac{-1+\sqrt{\mu}}{\mu-1} & \text{if } \mu \neq 1 \\ 1/2 & \text{if } \mu = 1\,. \end{cases}$$

This function is continuous and decreasing. Additionally, $f(0) = 1$ and $\lim_{\mu\to+\infty} f(\mu) = 0$. □

The interpretation of this result is that, as μ increases (where m_{21} increases relative to m_{12}), the threshold τ decreases, effectively biasing the decision in favor of c_1. In simpler terms, the cost influences the threshold to favor the less expensive option.

Remark 2

When comparing the threshold τ obtained through the MTh meta-algorithm with the conventional threshold in the binary case, $p^* = \frac{1}{1+\mu}$(as described in[27]), we find that the MTh threshold is not so optimal. Specifically:

$$\begin{cases} 0 < p^* < \tau < 1/2 & \text{if } m_{21} > m_{12} > 0 \\ 1/2 < \tau < p^* < 1 & \text{if } 0 < m_{21} < m_{12}\,. \end{cases}$$

In both cases, when the of misclassifying an instance as c_i is lower, the probability of assigning c_i increases with both methods. However, this increase is more pronounced with the classic thresholding method. The expression for p^* is derived by considering that the classic method assigns class c_1 if the expected misclassification cost is lower than assigning c_2, that is, if $m_{12}\,p_2 < m_{21}\,p_1$, resulting in $p_1 > 1/(1+\mu)$, taking into account that $p_2 = 1 - p_1$.

It is important to note that MTh offers the advantage of being readily applicable to multiclass scenarios and, in this context, satisfies the pseudo-optimality property outlined in Proposition 2. In simpler terms, MTh ensures that predicted class labels are not changed for instances if doing so would worsen the expected misclassification cost, as we demonstrate in the following result.

Proposition 2

In cases where predictions are made without ties, the MTh algorithm maintains the predicted class label, thereby avoiding a switch from c_k to c_ℓ, if the expected cost of misclassification as c_ℓ is not lower than that of misclassification as c_k. This pseudo-optimality feature ensures that decisions made by the algorithm do not lead to worse expected misclassification costs when choosing between class labels.

Proof

Let $(p_1, \ldots, p_r)$ denote the original a posteriori probabilities assigned to the class labels by any cost-insensitive classifier. The classifier's predicted class label is determined as:

$$c^* = c_k \quad \text{with } k = \arg\max_{i=1,\ldots,r} p_i\,.$$

After applying the MTh meta-algorithm, the chosen class label becomes $c^*_{th} = c_\ell$ if

$$\ell = \arg\max_{i=1,\ldots,r} \widetilde{p}_i = \arg\max_{i=1,\ldots,r} \frac{p_i}{\omega_i}\,.$$

Hence, the change in the predicted class label, i.e. $\ell \neq k$, occurs when:

$$p_k > p_\ell\,, \quad \text{but} \quad \frac{p_\ell}{\omega_\ell} > \frac{p_k}{\omega_k}\,.$$

This implies that $\omega_\ell < \omega_k$. In other words, the MTh meta-algorithm does not alter the predicted class label from the original $c^* = c_k$ to $c^*_{th} = c_\ell$ if the expected cost of misclassifying an instance as c_ℓ is not less than that of misclassifying it as c_k. □

Received: 14 February 2024; Accepted: 22 October 2024
Published online: 18 November 2024

References

1. Li, Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med.* **382**(13), 1199–1207. https://doi.org/10.1056/NEJMoa2001316 (2020).
2. Pallarès, N. et al. Characteristics and Outcomes by Ceiling of Care of Subjects Hospitalized with COVID-19 During Four Waves of the Pandemic in a Metropolitan Area: A Multicenter Cohort Study. *Infect Dis Ther.* **12**(1), 273–289. https://doi.org/10.1007/s40121-022-00705-w (2023).

3. Cester, A., Maselli, M. & Bolzetta, F. How to define the therapeutic limits. *Monaldi Arch Chest Dis.* **87**(2), 846. https://doi.org/10.4081/monaldi.2017.846 (2017).
4. Wang, A. Z. et al. Can we predict which COVID-19 patients will need transfer to intensive care within 24 hours of floor admission?. *Acad Emerg Med.* **28**(5), 511–518. https://doi.org/10.1111/acem.14245 (2021).
5. Zietz, M., Zucker, J. & Tatonetti, N. P. Associations between blood type and COVID-19 infection, intubation, and death. *Nat Commun.* **11**(1), 5761. https://doi.org/10.1038/s41467-020-19623-x (2020).
6. López-Otero, D. et al. Asociación entre el daño miocárdico y el pronóstico de pacientes hospitalizados por COVID-19, con y sin cardiopatía (in Spanish). *Registro CARDIOVID. Rev Esp Cardiol.* **74**(1), 105–108. https://doi.org/10.1016/j.recesp.2020.08.003 (2021).
7. Berenguer, J. et al. Development and validation of a prediction model for 30-day mortality in hospitalised patients with COVID-19: the COVID-19 SEIMC score. *Thorax* **76**(9), 920–929. https://doi.org/10.1136/thoraxjnl-2020-216001 (2021).
8. Berenguer, J. et al. Characteristics and predictors of death among 4035 consecutively hospitalized patients with COVID-19 in Spain. *Clinical Microbiology and Infection* **26**(11), 1525–1536. https://doi.org/10.1016/j.cmi.2020.07.024 (2020).
9. Lovejoy, C. A., Buch, V. & Maruthappu, M. Artificial intelligence in the intensive care unit. *Crit Care* **23**, 7. https://doi.org/10.1186/s13054-018-2301-9 (2019).
10. Chicco, D. & Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak.* **20**, 16. https://doi.org/10.1186/s12911-020-1023-5 (2020).
11. Wang, L., Lin, Z. Q. & Wong, A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci Rep* **10**, 19549. https://doi.org/10.1038/s41598-020-76550-z (2020).
12. Dalal, S., Singh, J. P., Tiwari, A. K. & Kumar, A. Identification of COVID-19 with CT scans using radiomics and DL-based features. *Netw Model Anal Health Inform Bioinforma* **13**(14). https://doi.org/10.1007/s13721-024-00448-3 (2024).
13. Ben-Gal, I. Bayesian Networks. In *Encyclopedia of Statistics in Quality and Reliability* (eds F. Ruggeri, R.S. Kenett & F.W. Faltin) (2008) https://doi.org/10.1002/9780470061572.eqr089
14. Neil, M., Fenton, N. & Tailor, M. Using Bayesian Networks to Model Expected and Unexpected Operational Losses. *Risk Anal.* **25**(4), 963–972. https://doi.org/10.1111/j.1539-6924.2005.00641.x (2005).
15. Delgado, R., González, J.L., Sotoca, A., & Tibau, X.A. A Bayesian Network Profiler for Wildfire Arsonists. In: Pardalos, P., Conca, P., Giuffrida, G., Nicosia, G. (eds) Machine Learning, Optimization, and Big Data. MOD 2016. *Lecture Notes in Computer Science* **10122**. Springer, Cham. (2016) https://doi.org/10.1007/978-3-319-51469-7_31
16. Delgado, R. & Sánchez-Delgado, H. Multi-instance learning with application to the profiling of multi-victim homicides. *Expert Systems with Applications* **237, Part B**, 121593. https://doi.org/10.1016/j.eswa.2023.121593 (2024).
17. Zhao, D. & Weng, C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *J Biomed Inform* **44**(5), 859–868. https://doi.org/10.1016/j.jbi.2011.05.004 (2011).
18. Delgado, R., Núñez-González, J.D., Yébenes, J.C., & Lavado, A. Vital Prognosis of Patients in Intensive Care Units Using an Ensemble of Bayesian Classifiers. In: Nicosia, G., Pardalos, P., Umeton, R., Giuffrida, G., Sciacca, V. (eds) Machine Learning, Optimization, and Data Science. LOD 2019. *Lecture Notes in Computer Science* **11943**. Springer, Cham. (2019) https://doi.org/10.1007/978-3-030-37599-7_51
19. Delgado, R., Núñez-González, J. D., Yébenes, J. C. & Lavado, A. Survival in the Intensive Care Unit: A prognosis model based on Bayesian classifiers. *Artificial Intelligence in Medicine* **115**, 102054. https://doi.org/10.1016/j.artmed.2021.102054 (2021).
20. McLachlan, S. et al. The fundamental limitations of COVID-19 contact tracing methods and how to resolve them with a Bayesian network approach. Preprint (2020) https://doi.org/10.13140/RG.2.2.27042.66243
21. Osarumwense, A.S., & Osayamen, O.K. A CoronaVirus Disease-2019 Prediction Model Based on Bayesian Belief Network. *International Journal of Academic Engineering Research (IJAER)* **4(4)**, 24–35 (2020) http://ijeais.org/wp-content/uploads/2020/4/IJAER200404.pdf
22. Fenton, N. A Note on UK Covid19 death rates by religion: which groups are most "at risk"? Preprint (2020) https://doi.org/10.48550/arXiv.2007.07083
23. Avila, E., Kahmann, A., Alho, C. & Dorn, M. Hemogram data as a tool for decision-making in COVID-19 management: applications to resource scarcity scenarios. *PeerJ.* **8**, e9482. https://doi.org/10.7717/peerj.9482 (2020).
24. Jian, C. et al. A pattern categorization of CT findings to predict outcome of COVID-19 pneumonia. *Front Public Health.* **8**, 567672. https://doi.org/10.3389/fpubh.2020.567672 (2020).
25. Abe, N., Zadrozny, B., & Langford, J. An Iterative Method for Multiclass Cost-Sensitive Learning. *Proc. 10th ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining*, 3–11 (2004) https://dl.acm.org/doi/pdf/10.1145/1014052.1014056
26. Zhou, Z. H. & Liu, X. Y. On multi-class cost-sensitive learning. *Computational Intelligence* **26**, 232–257. https://doi.org/10.1111/j.1467-8640.2010.00358.x (2010).
27. Elkan, C. The Foundations of Cost-Sensitive Learning. In Proceedings of the 17th International Joint Conference of Artificial Intelligence (IJCAI'01) **2**, 973–978. Seattle, Washington: Morgan Kaufmann (2001) https://dl.acm.org/doi/10.5555/1642194.1642224
28. Zadrozny, B., & Elkan, C. Learning and Making Decisions When Costs and Probabilities are Both Unknown. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 204–213 (2001) https://doi.org/10.1145/502512.502540
29. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**(1), 321–357. https://doi.org/10.5555/1622407.1622416 (2002).
30. Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. Effective prediction on three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In: 2016 IEEE International Conference of Online Analysis and Computing Sciences (ICOACS), Chongqing, China, 225–228 (2016) https://doi.org/10.1109/ICOACS.2016.7563084
31. Xu, Z., Shen, D., Nie, T. & Kou, Y. A hybrid sampling combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *J. Biomed. Inf.* **107**, 103465. https://doi.org/10.1016/j.jbi.2020.103465 (2020).
32. Shilaskar, S., Ghatol, A. & Chatur, P. Medical decision support systems for extremely imbalanced datasets. *Inf. Sci.* **384**, 205–219. https://doi.org/10.1016/j.ins.2016.08.077 (2017).
33. Delgado, R. & Núñez-González, J. D. Bayesian network-based Over-Sampling MEthod (BOSME) with application to indirect cost-sensitive learning. *Sci Rep.* **12**, 8724. https://doi.org/10.1038/s41598-022-12682-8 (2022).
34. Lomax, S. & Vadera, S. A cost-sensitive decision tree learning algorithm based on multi-armed bandit framework. *The Computer Journal* **60**(7), 941–956. https://doi.org/10.1093/comjnl/bxw015 (2017).
35. Ali, S. I. et al. Ensemble feature ranking for cost-based non-overlapping groups: a case study of chronic kidney disease diagnosis in developing countries. *In IEEE Acces* **8**, 215623–215648. https://doi.org/10.1109/ACCESS.2020.3040650 (2020).
36. Phankokkruad, M. Cost-sensitive extreme gradient boosting for imbalanced classification of breast cancer diagnosis. In 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 46–51 (2020) https://doi.org/10.1109/ICCSCE50387.2020.9204948
37. Mienye, I. D. & Sun, Y. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked* **25**, 100690. https://doi.org/10.1016/j.imu.2021.100690 (2021).
38. Domingos, P. MetaCost: A general method for making classifiers cost-sensitive. In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99). Association for Computing Machinery, New York, NY, USA, 155–164 (1999) https://doi.org/10.1145/312129.312220

39. Witten, I. H. & Frank, E. *Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations* (Morgan Kaufmann Publishers, 2005).
40. Xiaoyong, C., Deng, L., Yang, Q., & Ling, C.X. Test-Cost Sensitive Naïve Bayesian Classification. In Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04). Brighton, UK: IEEE Computer Society Press, 51–58 (2004) https://ieeexplore.ieee.org/document/1410266
41. Charlson, M., Szatrowski, T. P., Peterson, J. & Gold, J. Validation of a combined comorbidity index. *J. Clin. Epidemiol.* **47**(11), 1245–1251. https://doi.org/10.1016/0895-4356(94)90129-5 (1994).
42. Hruschka, E.R., Hruschka, E.R., & Ebecken, N.F.F. Feature Selection by Bayesian Networks. In: Tawfik, A.Y., Goodwin, S.D. (eds) Advances in Artificial Intelligence. Canadian AI 2004. *Lecture Notes in Computer Science* **3060**. Springer, Berlin, Heidelberg (2004) https://doi.org/10.1007/978-3-540-24840-8_26
43. Bielza, C. & Larrañaga, P. Discrete Bayesian Network Classifiers: A Survey. *ACM Computing Surveys* **47**(1), 1–43. https://doi.org/10.1145/2576868 (2014).
44. Scutari, M. Learning Bayesian Networks with the bnlearn R package. *Journal of Statistical Software* **35**(3), 1–22. https://doi.org/10.18637/jss.v035.i03 (2010).
45. Højsgaard, S. Graphical independence networks with the gRain package for R. *Journal of Statistical Software* **46**(10), 1–26. https://doi.org/10.18637/jss.v046.i10 (2012).
46. Team, R Core R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria (2022) https://www.R-project.org/
47. Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52**(3–4), 591–611. https://doi.org/10.2307/2333709 (1965).
48. "Student" Gosset, W.S. The probable error of a mean. *Biometrika* **6(1)**, 1–25 (1908) https://doi.org/10.2307/3001968
49. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* **1**(6), 80–83. https://doi.org/10.2307/3001968 (1945).

Acknowledgements

The authors would like to thanks their support to J. Baró, PhD at the CRM, for his initial work of cleaning with the dataset, and to the components of the MetroSud study group: Abelenda-Alonso, G., Rombauts, A., Rodríguez-Molinero, A., Gudiol, C., Aranda-Lobo, J., Arroyo, M., Pérez-López, C., Sanmartí, M., Moreno, E., Álvarez, M [a] C., Faura, A., González Bárcenas, M., Cruz Toro, P., Colom, M., Pérez, A., Serrano, L. They extend their sincere gratitude to the editor and reviewers for their valuable feedback and insightful suggestions, which have significantly contributed to the improvement of this manuscript. The author "Jordi Carratalà" is On behalf of MetroSud study group.

Author contributions

Generation and maintenance of the datasets: J. Carratalà, V. Diaz-Brito, E. Izquierdo, I. Oriol, A. Simonetti, S. Videla, C. Tebé, N. Pallarés **Dataset interpretation:** C. Tebé, N. Pallarés **Dataset cleaning and preprocessing:** R. Delgado, F. Fernández-Peláez **Methodology: conceptualization** R. Delgado **Methodology: implementation** R. Delgado **Writting:** R. Delgado, F. Fernández-Peláez

Funding

R. Delgado supported by Ministerio de Ciencia e Innovación, Gobierno de España, project ref. PID2021-123733NB-I00

Declarations

Conflict of interest

The authors declare no potential conflict of interests nor competing interests.github.com/RosDelgado/MTh

Code availability

The computer programming code (R function) that has been developed and utilized in this study, to implement Algorithm 1 (the Multi-Thresholding meta-algorithm, MTh) is accessible under an open-access (MIT) license. You can find it at https://github.com/RosDelgado/MTh.

Ethical implications

The use of predictive models in clinical settings raises several important ethical considerations. First and foremost, privacy and data security are critical: patient data used for training and validating models must be protected, and all practices must comply with data protection regulations. Second, it is essential to address potential biases within the model to ensure that predictions are equitable across different patient groups and to prevent discrimination in care. Third, predictive models should be used to complement, rather than replace, clinical judgment. It is important to be transparent about the limitations and uncertainties inherent in the model's predictions. Lastly, patients should be adequately informed about the use of predictive models in their care. Clear communication is necessary to help patients to understand how these models influence clinical decisions.

Additional information

Correspondence and requests for materials should be addressed to R.D.