Artificial Intelligence In Medicine 115 (2021) 102054

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Position Paper

Survival in the Intensive Care Unit: A prognosis model based on Bayesian classifiers

Rosario Delgado [a], J. David Núñez-González [a,b,*], J. Carlos Yébenes [c], Ángel Lavado [d]

[a] Department of Mathematics, Universitat Autònoma de Barcelona, Campus de la UAB, 08193 Cerdanyola del Vallès, Spain
[b] Department of Applied Mathematics, Engineering School of Gipuzkoa – Eibar Section, University of Basque Country (UPV/EHU), Otaola Av. 29, 20600 Eibar, Gipuzkoa, Spain
[c] Sepsis, Inflammation and Critical Patient Safety Research Group, Critical Care Department, Hospital de Mataró, Mataró, Spain
[d] Information Management Unit, Maresme Health Consortium, Hospital de Mataró, Mataró, Spain

ARTICLE INFO

Keywords:
Intensive Care Unit
Mortality risk
Bayesian classifier ensemble
Area Under the Curve
F-score
APACHE II

ABSTRACT

We develop a predictive prognosis model to support medical experts in their clinical decision-making process in Intensive Care Units (ICUs) (a) to enhance early mortality prediction, (b) to make more efficient medical decisions about patients at higher risk, and (c) to evaluate the effectiveness of new treatments or detect changes in clinical practice. It is a *machine learning* hierarchical model based on Bayesian classifiers built from some recorded features of a real-world ICU cohort, to bring about the assessment of the risk of mortality, also predicting destination at ICU discharge if the patient survives, or the cause of death otherwise, constructed as an ensemble of five base Bayesian classifiers by using the average ensemble criterion with weights, and we name it the Ensemble Weighted Average (EWA).

We compare EWA against other state-of-the-art *machine learning* predictive models. Our results show that EWA outperforms its competitors, presenting in addition the advantage over the ensemble using the majority vote criterion of allowing to associate a confidence level to the provided predictions. We also prove the convenience of locally recalibrate from data the standard model used to predict the mortality risk based on the APACHE II score, although as a predictive model it is weaker than the other.

1. Introduction

Medical care is one of the most exciting frontiers in data mining and *machine learning*. Although the methodology of prognostic research, including the prediction rules and the approaches to validate them, is still relatively underdeveloped, accurate *prognosis*, which refers to a prediction of the course and outcomes of a patient based on the most likely trajectory of a disease or health problem, has become a key concept in patient care today.

Clinical decision-making for critically ill patients admitted in Intensive Care Units (ICU) is a costly and complex process, which suffers from excessive variability between the opinion of physicians, since it is largely driven by experience and instinct [1,2]. Apart from age, comorbidities or organ failures, there are other aspects related to he death of patients in ICU, such as delay on attention or inadequate management, which are also linked to the length of stay and costs, as well as to the decrease in quality of life at ICU discharge in survivors [3–5]. In order to improve the quality of the attention, it is important to establish protocols for the management of the healthcare process [6,7].

The traditional approach to improve the performance of ICUs is founded on the development of scores which try to predict the likelihood of negative outcomes (e.g. risk of death). From them, the Acute Physiology And Chronic Health Evaluation (APACHE) is a commonly used scoring system to quantify the severity of illness and to group adult ICU patients by predicted risk of mortality, based on patient data corresponding to the first 24 h after admission to the ICU. This prediction is carried out by means of a logistic regression model in which APACHE is one of the regressors, validated on previous groups of ICU patients [8].

* Corresponding author at: Department of Mathematics, Universitat Autònoma de Barcelona, Campus de la UAB, 08193 Cerdanyola del Vallès, Spain and Department of Applied Mathematics, Engineering School of Gipuzkoa – Eibar Section, University of Basque Country (UPV/EHU), Otaola Av. 29, 20600 Eibar, Gipuzkoa, Spain.
E-mail addresses: delgado@mat.uab.cat (R. Delgado), josedavid.nunez@ehu.eus (J.D. Núñez-González), jyebenes@csdm.cat (J.C. Yébenes), alavado@csdm.cat (Á. Lavado).

https://doi.org/10.1016/j.artmed.2021.102054
Received 30 October 2020; Received in revised form 1 March 2021; Accepted 17 March 2021
Available online 23 March 2021

Although there are different versions of this score, APACHE II [9] is still the most used today (see for example the recent works [10], [11] and [12]). This traditional approach based on APACHE II and its successive versions presents some limitations, such as:

(a) not incorporating variations between units or regions,
(b) better behaviour in large populations than in small ones, and
(c) stiffness as predictive model, since if the value of any of the key variables for the patient is not known (variables related with the kind of admission and the severity score APACHE itself), cannot provide a prediction,

which make it unsatisfactory. With its more than 40 years, in some sense APACHE score has become obsolete, taking into account the evolution of medical practice today. In addition, the impact of age on survival, one of the items that scores the most on it, has changed, as well as the life expectancy of neoplastic, coronary or HIV patients, for example. APACHE has not adapted to this new reality.

Thus, there is a need to find out new methods to address these shortcomings and to improve the predictions of the risk of mortality for ICU patients. The application of Artificial Intelligence to Medicine to build predictive prognostic models may represent an opportunity for improvement over scale-based models, such as the one that uses the APACHE score. It is our purpose in this work to present a machine learning methodology that will be validated with a real database. Experimentally we see that it provides good results and avoids the weak points of the traditional approach, so it proves to be a better alternative to the regression models based on the APACHE score.

1.1. Literature review

Different data-driven models have been considered in the literature to support medical experts in their clinical decision-making process in hospital ICUs, to improve quality and for benchmarking purposes on the one hand, and for a greater personalization of care on the other. These models could reduce the inter-clinical variability and are able of treating a large number of variables, finding complex relationships between them. In recent years we can find several works on the state-of-the-art of the use of quantitative methods to assist in medical prognosis and decision-making in health centers and hospitals [13–15]. Some of these methods use *machine learning* tools to face with different situations; in particular, there have been several attempts to apply *machine learning* to improve the management of ICUs (see [16–20]).

Examples of recent works are: [21], in which the authors introduce a predictive model for the survival probability via support vector machines (SVM), making comparisons with other based on logistic regression (LR), where the APACHE score is recalibrated, and they show that SVM outperforms others. Benchmark results for mortality and length of stay predictions using Deep Learning have been presented in [22], where an ensemble of *machine learning* models and some scores have also been considered, using the Medical Information Mart for Intensive Care III (MIMIC-III) publicly available dataset [23], and showing that Deep Learning models consistently outperform the other approaches. To predict the risk of death from some quantitative measures based on the heart rate signals of ICU patients suffering cardiovascular diseases, eight supervised classifiers have been introduced in [24] with the MIMIC-III dataset: decision tree, linear discriminant, logistic regression, SVM, random forest (RF), boosted trees, Gaussian SVM, and K-nearest neighbors (K-NN), showing that the former performs better than the others. A deep multi-scale convolutional architecture trained on the MIMIC-III dataset for mortality prediction has been introduced in [25], to address the problem that although deep neural networks are able to outperform the score approach, they suffer from lack of interpretability. The same problem has been considered in [26], proposing a different solution: an interpretable Bayesian neural network architecture which offers the flexibility of neural networks without sacrificing interpretability in terms of the selected features, and that has been evaluated using two real-world ICU datasets, MIMIC-III and CENTER-TBI [27]. We finish this review on related works with [28] and [29], which use multiple *machine learning* methods to improve prediction performance.

1.2. Black versus white box models in pattern recognition

Pattern recognition consists of classifying objects or individuals, which are described by a set of characteristics or features, using a model built on the basis of some data, assigning them a class label. In the *supervised learning* pattern recognition problems, each object or individual in the data comes with an observed label, and all the information relating to an object or individual forms a "case". Then, the task of pattern recognition is to construct (train) a model, that is, a *classifier*, to assign a class to each new case.

Neural networks (NN), which are designed to mimic the performance of the human brain, is by far the most widely used *machine learning* methodology for pattern recognition in the environment of an hospital ICU so far. Its major weakness, however, is that the procedure by which NN discovers relationships or patterns in the data is hidden or opaque (it is said that NN is a *black box*) and therefore it is not easily understood nor explained. That is, although the predictive capacity of this classifier usually shows its adequacy in practice, the classification knowledge learned by the NN turns out to be obscure.

Instead, using Bayesian Networks (BN) is a good way to model situations under uncertainty since, unlike what happens with the *black box* models, they are characterized by being *white box* models that show the relationships and patterns found between the variables in a completely understandable (transparent) way. For this reason BN have been gaining popularity as classifiers in health care applications, thanks to their versatility and power. Just to mention some examples, they have been used in public health evaluation [30], for risk assessment with emerging diseases [31] or for medical diagnosis [32]. BN have been used in the Intensive Care Unit to evaluate EEGs [33] and to establish prognosis in patients with head injuries [34]. To finish this brief summary, the application of Dynamic Bayesian network has been useful for the prediction of organic failure sequences in patients admitted to the ICU in [35].

The authors of [36] use Naive Bayes, jointly with other *machine learning* methodologies, as predictive tools for the inference of lactate level and mortality risk regarding sepsis. Just one comment on this: although Naive Bayes is a particular case of BN, it is not a *white box* model but a *black box*, since its structure (DAG) is fixed and not learned from the data, so it does not reflect the dependency relationships between the variables included in the model. Naive Bayes is based on a very strong independence assumption between the features conditioned to the class variable; despite that, it has shown to work quite well in many complex real-world situations.

1.3. The methodology

With the aim of helping in the vital prognosis of patients admitted to the ICU of a hospital, we propose a machine learning methodology for pattern recognition purpose, consisting in the use of ensembles of BN to build a hierarchical predictive prognosis model.

The idea behind the *ensembles of classifiers* is to combine several individual classifiers to get a new one that beats them all. It seems a natural strategy since we tend to seek input from different people before making our important decisions, and this is especially so in the field of clinical diagnosis, where the opinions of different experts can be taken into account to reach the final decision about a patient. Instead of putting the emphasis on choosing a good classifier, if there is one, we put it on the combination of various classifiers, in the hope that by combining them, the faults of some will be compensated by the others, and the joint result improves each of the parts.

The prediction process with the hierarchical model consists of two stages: (1) predicting the class variable *Result* (live/die), (2) predicting the class variables *Destination* (at ICU discharge) or *Cause* (of death), depending on the prediction in the previous stage. We consider the cause of death as an essential element in the hierarchical predictive prognostic model, whose prediction can help to improve the evaluation of the quality of the care process at the ICU level. Both stages lean on an ensemble of five base Bayesian classifiers, that we denote by EWA, constructed using the *weighted average* criterion with appropriate weights. This criterion is of the type *"fusion of continuous-valued outputs"*, that unlike what happens with the criteria based on the *"fusion of labels"*, such as the *majority vote*, is based not on the prediction, but on the probabilities assigned to the classes by each of the classifiers that make up the ensemble, and has the advantage of being compatible with the MAP (Maximum A Posteriori) criterion. About the weights, our proposal is to use an adequate transformation of the Area Under the Precision-Recall curve (AUPR), which is used instead of Accuracy as performance metric since our dataset is quite skewed for the class variable.

Up to our knowledge, this is a novel approach for vital prognosis of critically ill patients, both by the fact of using a hierarchical model, and by the use of an ensemble of BN as machine learning methodology for pattern recognition, which is based on the weighted average criterion with weights defined from the AUPR with an adequate transformation.

1.4. Evaluation

We apply the proposed methodology to a real-world dataset of critical patients admitted to the ICU of a hospital. To show its usefulness to aid in the vital prognosis, we carried out an experimental evaluation to make comparisons with other pattern recognition proposals, since the performance of the prediction models appears to be, usually, context dependent.

For that, we compare EWA with other state-of-the-art *machine learning* methodologies, such as NN, SVM and RF, without intending to be exhaustive, but rather to highlight its strengths and weaknesses. We also compare the standard mortality prediction based on the APACHE II score using a logistic regression against a locally recalibrated model that we also built using APACHE II, for which coefficients are estimated from the data by means of a logistic regression, and then compare them both with the EWA ensemble. In addition, we compare EWA with the ensemble but without weights, denoted by EA (or, to be more rigorous, with weights all the same), and also with the ensembles obtained from the same base Bayesian classifiers by using both the majority vote criterion (denoted by MV) and the weighted majority vote criterion (denoted by WMV), that have already been introduced in our proceedings paper [37].

Both from the medical and from the management of the ICU points of view, the hierarchical model based on the EWA ensemble gave interesting results from where we get relevant conclusions. Besides, it allows to associate a reasonable confidence level to predictions, issue on which the ensembles MV and WMV fail. Moreover, we have implemented this model as inference engine of an expert system that helps in the vital prognosis at the ICU level, and developed a computational tool intended to make easier the communication between the medical staff and the expert system.

1.5. A further utility

In each specific context it may happen that some of the features might be irrelevant for the prognosis of patients (they are only "noise"), while others show to be important for prediction, and even some features might be important but only in relation to others. That is, they are not all equally relevant.

As said in Section 1.2, one of the advantages of our predictive prognosis model is that it is a white box and, therefore, allows us to to carry out a study (see Section 4) on the importance of the features that are included in the model based on the consideration of two different aspects: *centrality and betweenness* and *feature strength*. Centrality and betweenness, on the one hand, are concepts of the field of Graph Theory and Network Analysis that can be applied to BN to identify the most "influential" variables in the model, in a sense that will be specified. On the other hand, we introduce a measure of the feature strength based on a statistical distance between the *a posteriori* conditional probability distributions of the class variable given different values of a fixed feature. In Section 4 we also consider the odds ratio (OR), which is a well known quantification of the strength of the association between two events, that in our context will be "die" when a fixed feature is present, and when it is absent.

The organization of the rest of the paper is as follows: in Section 2 we present the data set and the hierarchical model that we will use to

Table 1
List of variables (part I).

1. Demographic characteristics	% respect to non-missing values
F_1 : Sex	
Male	63.6%
Female	36.4%
F_2 : Age	Median: 70
	Q_1, Q_3: 59, 7
Ranges:	
<45	8.8%
45–54	9.8%
55–64	19.4%
65–74	24.9%
75–84	26.2%
> 84	10.9%
2. Comorbidities	
F_3 : Charlson comorbidity index	
0	31.7%
1	24.2%
2	15.9%
3	10.5%
>3	17.7%
3. Admission	
F_{17} : Origin (location before ICU admission)	
Ward	20.2%
Operation Room	14.0%
Emergency Room	41.0%
Extra Hospital Emergency	1.7%
Other Hospital	23.1%
F_{18} : Generic syndrome (causing admission)	
Elective Surgical	6.5%
Urgent Surgical	9.8%
Coronary	17.5%
Medical	64.3%
Trauma	1.9%
F_{19} : Sepsis (at admission)	
Yes	35.7%
No	64.3%
Main cause of admission (yes/no)	% of yes
F_4: ACS (Acute Coronary Syndrome)	18.7%
F_5: RF (Respiratory Failure)	33.0%
F_6: Shock	27.1%
F_7: Coma	7.3%
F_8: Renal F (Renal Failure)	4.1%
F_9: Hepatic F (Hepatic Failure)	0.2%
F_{10}: CRA (Cardio Respiratory Arrest)	4.8%
F_{11}: ES (Elective Surgical)	6.7%
F_{12}: Arrhythmia	4.1%
F_{13}: CT (Cranial Trauma)	0.2%
F_{14}: OT (Other Trauma)	1.3%
F_{15}: Intoxication	1.0%
F_{16}: Other syndromes	6.3%

Table 2
List of variables (part II). *Destination* = "Morgue" if *Result* = "die". *Cause* = "Not Dead" if *Result* = "live". We have merged classes "Septic Complications" and "Non-septic Complications" (1.57% and 1.53%, respectively) for variable *Cause*, into the single class "Complications".

4. Severity (on first 24 h of admission)	
F_{20} : ICU workload (therapeutic requirements)	
Medical monitoring	25.4%
Medical unstable with coma or shock	22.4%
Medical unstable without coma neither shock	21.2%
Post-surgical monitoring	5.1%
Post-surgical unstable	25.9%
F_{21} : APACHE II	Median: 13
	Q_1, Q_3: 8, 18.25
Ranges:	
<5	9.0%
5–9	25.6%
10–14	23.6%
15–19	19.6%
20–24	11.5%
25–29	6.2%
30–34	2.7%
>34	1.8%
Outcomes	
Result	
live	85.3%
die	14.7%
Destination (at ICU discharge)	
First Attention Hospital	77.7%
Major Complexity Hospital	7.6%
Morgue	14.7%
Cause (of death)	
Cause of Admission	11.2%
Complications	3.1%
Not Dead	85.7%

predict the risk of mortality in the ICU, as well as the destination for those patients who are expected to survive, or the cause of death for the rest. Specifically, we introduce Bayesian networks, used as base models, and the ensembling of classifiers. We also explain the standard method used to predict the risk of mortality at the ICU level based on the APACHE II score, as well as the implementation and validation procedures. Section 3 shows the results we have obtained, Section 5 is devoted to the final comments and conclusions, and the appendices include some figures and tables.

2. Materials and methods

2.1. Dataset description

Our dataset is a cohort of 2510 critical patients admitted to the ICU of the Mataró Hospital (Mataró, Spain) from years 2016 (661 patients), 2017 (693), 2018 (663) and 2019 (493). With the aim of predicting mortality/survival at the ICU first, and then the destination at ICU discharge for patients who survive their stay, or the cause of death for patients who pass away, different features of the patients have been considered (see Tables 1 and 2). ICUs can be thematic (related to a specific kind of patient, as can be neuro-trauma ICU, Coronary unit, medical ICU, post-surgical ICU, …) or polyvalent, as in our case. To clarify the syndromic classification of critically ill patients, as is usual in polyvalent ICUs, we use four categories:

1. Demografic characteristics
   Sex (F_1)
   Age (F_2)
2. Comorbidities (Charlson comorbidity index, F_3)
3. Admission
   Origin (location of patient before ICU Admission, F_{17})
   Generic syndrome (that cause admission, F_{18})
   Sepsis (F_{19})
   Main cause of admission (F_4–F_{16})
4. Severity (on first 24 h of admission)
   ICU workload (therapeutic requirements, F_{20})
   APACHE II score (F_{21})

In general, on admission we classify critically ill patients attending to the generic syndrome (F_{18}) into

–Surgical (a major invasive procedure is related to the cause or the treatment of admission). We distinguish between "elective" and "urgent".
–Coronary (admission related to a coronary syndrome).
–Medical (no acute coronary syndrome neither major invasive procedures related to the cause or treatment of admission).
–Trauma (in case of physical external agent damage).

Related to severity on first 24 h of admission, the therapeutic requirements (ICU workload F_{20}) of medical (including coronary) or surgical (including trauma) patients, depending on the presence or not of organ failure, can be

	Medical patient	Surgical patient
Presence of organ failure	Medical unstable	Medical monitoring
Absence of organ failure	Post-surgical unstable	Post-surgical monitoring

Stable patients without organ failure just require monitoring to prevent complications, while unstable patients require specific organ failure support, and in this case, we distinguish if they are (or not) in coma or shock. Patients in coma (if Glasgow Coma Score is under 9) or shock (requirement of vasoactive agents to maintain organ perfusion) present the highest mortality and require the highest therapeutic effort.

The presence or absence of sepsis at admission (F_{19}) allows a better understanding of the patient's characteristics. But this is a too nonspecific classification, so we can make an additional classification according to more specific syndromes grouped in the category of "Main cause of admission" (see Table 1). Note that despite that some syndromes can overlap in the same patient, we only identify the most severe condition that causes the ICU admission. For example, in case of a patient in coma, due to a shock secondary to a pancreatitis infarction, the primary specific syndrome is shock (F_6 = "yes"), the generic syndrome is F_{18} = "Medical", without sepsis (F_{19} = "no") and the ICU workload is F_{20} = "Medical unstable with coma or shock".

Although some variables (F_9, F_{13}, F_{14}, F_{15}) or categories (F_{17} = "Extra Hospital Emergency", F_{18} = "Trauma") have very little presence in the current cohort, they have been kept in the study to be able to explore all the possibilities of patient flow, when the database grows.

All variables were, or have been transformed into type factor through a discretization procedure. Age variable has been categorized as well as is done with APACHE II score. In the same way, variable "Charlson comorbidity index", that can show integer values from 0 to 29, has been discretizated into 5 categories: 0, 1, 2, 3 and >3. Missing values are infrequent, and only appear in 9 of the 24 variables, none of them of the "Main cause of admission" category. As expected, patients with missing values in variable *Result* also have missing values in *Destination* (at ICU discharge) if the variable *Result* is *live* and in *Cause* (of death) for patients for which the variable *Result* is *die*.

We observe that 63.6% of patients are male, with a median age of 70 years (average of 67.34) and that mortality at the ICU is 14.7%. "First attention hospital" is the destination at ICU discharge for 91% of patients that survive, and among the patients that did not survive, the cause of death of the 78.4% was the cause of admission, and only for a 21.6% it was a complication suffered at the ICU, of which, half are of a

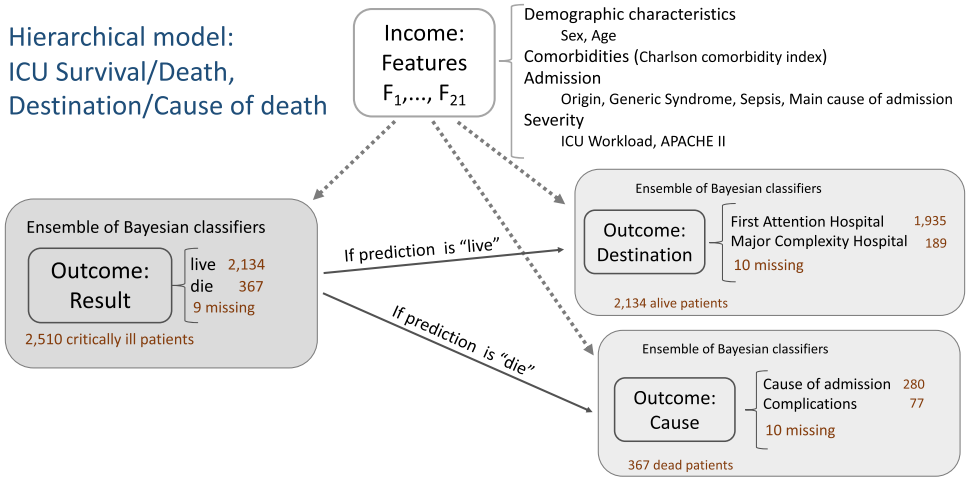

Fig. 1. Processing pipeline of vital prognosis (survival/mortality) prediction for patients in Intensive Care Units.

septic nature.

In Fig. 2 in Appendix A we observe the temporal evolution of the rate of mortality at the ICU, both for the overall population and disaggregated by sex. Instead, in Fig. 3 we see the evolution of the mortality rate with age, also disaggregated by sex and for the overall population. We can observe that the mortality rate is greater for females (except for the youngest patients), and that it increases form young to older people but decreases for the oldest. Finally, in Figs. 4 and 5 we show the distribution of the missing values. The colors in Fig. 5, from white to black in the gray scale, correspond to the different categories that each of the variables takes, while the red color is reserved to indicate missing values. We do not observe any pattern in the distribution of missing values among cases, and as for the variables, those that present missing values are those of the "Admission" and "Severity (on first 24 h of admission)" categories, especially F_{21}: APACHE II, with a 12.67%.

2.2. Building the hierarchical model

The hierarchical model consist of three parts, namely, the classifier for predicting the variable *Result* (live/die), at the first stage, and two more classifiers at the second stage, one for predicting the variable *Destination*, and the other for predicting *Cause*, depending on whether the prediction in the first stage was "live" or "die", respectively, as can be seen in Fig. 1.

Each of the parts consists of an ensemble of BN built using the *weighted average* criterion with appropriate weights, denote by EWA. Next we explain what Bayesian networks are and how ensembles are constructed.

2.2.1. Bayesian networks

Bayesian networks (BN) are graphical models representing the probabilistic relationships among variables affecting a phenomenon, which are used for probabilistic inference. For a set of random variables, a BN is a model that represents their joint probability distribution P, the graphical part of the model consisting of a *directed acyclic graph (DAG)*, whose nodes represent the random variables. The directed arcs among the nodes represent conditional dependencies (not necessarily causal) governed by the **Markov condition**, which establishes that each node in the DAG is independent of those who are not its descendants given its parents are known. When a Bayesian network is used to classify cases into a set of categories or classes, we term it *Bayesian classifier*.

(Bayesian) inference is the term used to refer to the update of probabilities of the network from a given evidence: we compute *a posteriori* probabilities from evidences and *a priori* probabilities. Prediction of a query variable X given the evidence E is the instantiation of X with the largest *a posteriori* probability, and this probability is said to be the *confidence level* of the prediction.

Table 3
Traits of the five base Bayesian classifiers used to construct the ensembles.

Classifier	Score	Restriction on the directed arcs
BC_1 (Naive)		Whitelist: from class to each feature. Blacklist: among features
BC_2	BIC	Whitelist: from class to each feature
BC_3	AIC	Blacklist: from each feature to class
BC_4	AIC	Whitelist: from class to each feature
BC_5 (TAN)		Whitelist: from class to each feature Each feature has an extra incoming arc from other feature

To predict the risk of death of critically ill patients, we learn five different base Bayesian classifiers from data, say BC_1, …, BC_5, by considering the features and the class variable *Result*. This allows further enquiry into the relationships between the features and the vital prognosis, being this an advantage over typically *black box machine learning* methods, such as Neural Networks, which are unable to provide explanations for their predictions. In Table 3 we report the traits of construction of these classifiers, including the score function for the structure learning (learning of the DAG) and the restrictions on the allowed directed arcs, in the form of *whitelist/blacklist* of forced/forbidden arcs. Maximum Likelihood Estimation is used to estimate the parameters.

Naive Bayes has a fixed structure (DAG) which is not learned from the data, and assumes that features are independent of each other given the class, which can be unrealistic in many applications. The other four classifiers in Table 3 are different attempts to improve classification by relaxing this assumption and trying, at the same time, to maintain simplicity and efficiency as much as possible. In particular, TAN (Tree Augmented Naive) relaxes the feature independence assumption of the Naive Bayes through a tree structure, in which each feature only depends on the class and one other feature. Note that both, BC_2 and BC_4 are Augmented Naive Bayes classifiers [39] since the class variable is assumed to be a root node parent of every feature, and the subgraph of the features is an unrestricted Bayesian network.

2.2.2. Ensembles based on the fusion of labels outputs

In [37] we built an ensemble of classifiers to predict the class variable *Result*, say WMV, acronym for Weighted Majority Vote (denoted by EBC there), from the five base classifiers BC_1, …, BC_5, with the *weighted majority vote* criterion, which is a single-winner voting system but in which more power is given to more "competent" base classifiers. This criterion, as the *majority vote*, falls into the *fusion of labels outputs* ensemble methods. Concretely, fixed a critical patient and a class j, we

Table 4
Toy example 1. Confidence level for the prediction given by MV is <0.5.

Classifier	Weigths	Prob. of "die"	Prediction	Pred. MV	Pred. WMV
BC_1	$w_1 = 0.25$	$p_1 = 0.55$	die		
BC_2	$w_2 = 0.10$	$p_2 = 0.55$	die		
BC_3	$w_3 = 0.05$	$p_3 = 0.55$	die	die (0.24731<0.5)	live (0.66236>0.5)
BC_4	$w_4 = 0.30$	$p_4 = 0.10$	live		
BC_5	$w_5 = 0.30$	$p_5 = 0.10$	live		

Table 5
Toy example 2. Confidence level for the prediction given by WMV is <0.5.

Classifier	Weights	Prob. of "die"	Prediction	Pred. MV	Pred. WMV
BC_1	$w_1 = 0.25$	$p_1 = 0.95$	die		
BC_2	$w_2 = 0.10$	$p_2 = 0.95$	die		
BC_3	$w_3 = 0.05$	$p_3 = 0.95$	die	die (0.95324>0.5)	live (0.32725<0.5)
BC_4	$w_4 = 0.30$	$p_4 = 0.45$	live		
BC_5	$w_5 = 0.30$	$p_5 = 0.45$	live		

consider the **discriminant** function $D_j = \sum_{i=1}^{5} w_i d_{i,j}$ where $d_{i,j} = 1$ if classifier i assigns class j to the patient, and 0 otherwise, and $w_i, i = 1, \ldots, 5$ are the weights of the five base classifiers, that is, D_j is the sum of weights corresponding to classifiers that assign the patient to class j. The inferred class for the given patient by the WMV classifier is taken to be the one that maximizes the discriminant function. (Note that with $w_i = w_j$ for all $i, j = 1, \ldots, 5$, this rule corresponds to the mere criterion of the *majority vote* and we denote by the acronym MV the corresponding ensemble.)

For the assignment of weights to the base classifiers, and bearing in mind that the combination of unbalanced data (14.7% "die" in variable *Result*) and a small sample size (2, 510 patients) prevents the use of *Accuracy* as an evaluation metric in classification, we followed [40] and [41] when considering a measure based on the *Recall* (also called *Sensitivity*) and the *Precision*, with "positive class" the minority class *die*, which provides a good representation of performance assessment in the binary classification: the *Area Under the Precision-Recall curve* (AUPR), being the *Precision-Recall* (PR) curve that obtained by plotting *Precision* over *Recall*. The PR curve provides a more informative picture of the performance of the classifier than the Receiver Operator Characteristic (ROC) curve when dealing with highly skewed datasets, as is our case. For example, in [42] AUPR has been used for mortality and decompensation tasks since the MIMIC-III dataset, which is the one used by the authors for experimentation purposes, suffers from class imbalance. Considering the above, we assign a weight w_i to the base classifier i, which is obtained from its estimated AUPR, denoted by $A_i \in [0, 1]$, in the following way:

$$w_i = \frac{h_i}{\sum_{j=1}^{5} h_j}, \quad \text{where} \quad h_i = log\left(\frac{\frac{1}{2}(A_i + 1)}{1 - \frac{1}{2}(A_i + 1)}\right) \tag{1}$$

Note that

$$\frac{1}{2}(A_i + 1) \in [0.5, 1],$$

and that therefore,

$$\frac{\frac{1}{2}(A_i + 1)}{1 - \frac{1}{2}(A_i + 1)} \geq 1$$

and consequently $h_i \geq 0$. This transformation of the A_i's is a dilatation since if $A_i < A_j$, therefore $h_j - h_i > A_j - A_i > 0$. With this assignment of weights, we magnify the relevance of the base classifiers using weights based on the AUPR metric.

The WMV classifier proved to have a good performance in [37] but it presents the problem of not following the criterion of **maximum a posteriori probability (MAP)**, which is optimal in the sense of minimizing the expected 0–1 loss. For this reason, it is perfectly possible that the predicted class in the binary case may have an associated confidence level less than 0.5, what is counterintuitive and difficult to justify. Let us show two toy examples in Tables 4 and 5 , exemplifying this paradoxical situation, with classes "die" ($j = 1$) and "live" ($j = 2$), and the confidence levels in brackets. In both examples, weights are the same and the prediction of any base classifier as well, so the MV classifier gives 3 votes to "die" class and 2 votes for "live" class, that is,

$$D_1 = 3 > D_2 = 2,$$

bringing us to "die" as prediction, while for the WMV,

$$D_1 = \omega_1 + \omega_2 + \omega_3 = 0.4 < D_2 = \omega_4 + \omega_5 = 0.6,$$

resulting from this that the prediction is "live", which is the opposite of the class predicted with the MV. What changes from one example to the other is the probability of "die" of the base classifiers. We can observe that in the first example, the confidence level associated to the prediction provided by the MV is <0.5, while the same happens for the WMV in the second example.

The confidence levels in Tables 4 and 5 have been computed in the following way: for the MV, the confidence level is the probability that the majority of the votes (3, 4 or 5) be for "die", that is:

$$\mathrm{CL_{MV}} = \prod_{\ell=1}^{5} p_\ell + \sum_{j=1}^{5}((1-p_j)\prod_{\substack{\ell=1\\ \ell\neq j}}^{5} p_\ell) + \sum_{j=1}^{5}\sum_{\substack{k=1\\ k\neq j}}^{5}((1-p_j)(1-p_k)\prod_{\substack{\ell=1\\ \ell\neq j,k}}^{5} p_\ell),$$

with p_ℓ being the probability of "die" for the ℓth base classifier. For the WMV, the confidence levels have been computed as the probability that the sum of the weights of the base classifiers that vote "die" be >0.5, which is dependent on weights, by means of the following formula (note that in this case ties are not possible since there is no combination of weights whose sum is exactly 0.5):

$$\mathrm{CL_{WMV}} = \prod_{\ell=1}^{5} p_\ell + \sum_{r=1}^{4} \sum_{\substack{i_1,\ldots,i_r=1\\ (i_1,\ldots,i_r)\in\Delta^r_{w_1,\ldots,w_5}}}^{5} \left(\prod_{h=1}^{r}(1-p_{i_h}) \prod_{\substack{\ell=1\\ \ell\neq i_1,\ldots,i_r}}^{5} p_\ell\right)$$

Table 6
Toy example 3. Predictions with EA and EWA are different (confidence levels are in brackets).

Classifier	Weights	Prob. of "die"	Prob. of "live"	Prediction	Pred. EA	Pred. EWA
BC_1	$w_1 = 0.25$	$p_1 = 0.90$	$1 - p_1 = 0.10$	die		
BC_2	$w_2 = 0.10$	$p_2 = 0.90$	$1 - p_2 = 0.10$	die		
BC_3	$w_3 = 0.05$	$p_3 = 0.90$	$1 - p_3 = 0.10$	die	die (0.60>0.5)	live (0.55>0.5)
BC_4	$w_4 = 0.30$	$p_4 = 0.15$	$1 - p_4 = 0.85$	live		
BC_5	$w_5 = 0.30$	$p_5 = 0.15$	$1 - p_5 = 0.85$	live		
					$\widetilde{D}_1 = 0.60$	$\widetilde{D}_1 = 0.45$
					$\widetilde{D}_2 = 0.40$	$\widetilde{D}_2 = 0.55$

where

$$\Delta^r_{w_1,\dots,w_5} = \{(i_1,\dots,i_r) : 1 \le i_1,\dots,i_r \le 5, i_1 \ne \cdots \ne i_r, \sum_{\ell=1}^{r} w_{i_\ell} < 0.5\}.$$

2.2.3. Ensembles based on the fusion of continuous-valued outputs

In this paper we consider, as the main novelty regarding [37], two ensembles based on combiners that fall into the *fusion of continuous-valued outputs*, which are the simple mean (average), denoted by EA hereinafter, and the weighted average with the weights given by (1), denoted by EWA from now on. More specifically, fixed a critical patient and a class j, let us introduce the **discriminant** function $\widetilde{D}_j = \sum_{i=1}^{5} w_i \widetilde{d}_{i,j}$ where $\widetilde{d}_{i,j}$ is the probability that classifier i assigns the class j to the patient, and weights w_i are given by (1).

The inferred class for the given critical patient by the EWA classifier is taken to be the one that maximizes the discriminant function $\widetilde{D}$. Therefore, with this criterion, the confidence level associated to class j is $\widetilde{D}_j$ since $\widetilde{D}_1 + \widetilde{D}_2 = 1$, which implies compatibility with the **MAP criterion**. (Note that with $w_i = w_j = 1/5$ for all i, $j = 1$, …, 5, this rule corresponds to the simple mean combiner EA.) In the toy examples 1 and 2 (Tables 4 and 5, respectively) we can apply both the EA and the EWA, and obtain the respective predictions and confidence levels. Indeed, for the EA in Table 4,

$$\widetilde{D}_1 = \sum_{i=1}^{5} \frac{p_i}{5} = 0.37 < \widetilde{D}_2 = \sum_{i=1}^{5} \frac{1 - p_i}{5} = 0.63$$

while for the EWA,

$$\widetilde{D}_1 = \sum_{i=1}^{5} \omega_i p_i = 0.28 < \widetilde{D}_2 = \sum_{i=1}^{5} \omega_i (1 - p_i) = 0.72$$

which bring us to the "live" prediction with both classifiers, and respective confidence levels of 0.63 and 0.72, both >0.5. Analogously, for the toy example in Table 5, both ensembles give as prediction "die", with respective confidence levels of 0.75 and 0.65, both >0.5. Although in these two toy examples the predictions of EA and EWA have coincided with each other, this does not necessarily have to happen in general. For instance, consider the third toy example in Table 6.

2.3. Implementation

The processing pipeline of the vital prognosis is summarized in Fig. 1. After survival/mortality prediction, outcome variable *Destination* will be predicted in a second step, if the prediction for *Result* is "live", by using an appropriate classifier which will be an ensemble similar to that used for predicting *Result* but substituting the response variable *Result* for *Dest*. Otherwise, if the prediction for *Result* is "die", another ensemble similar to that used for *Result* but substituting *Result* as outcome for *Cause*, will be used to predict the cause of death.

Learning and prediction algorithms have been implemented in R language. For structure learning of the base Bayesian networks BC_2, BC_3 and BC_4, hill-climbing score-based structure learning algorithm has been used, implemented by function **hc** of the *bnlearn* package [43], whereas for BC_5 the **tree.bayes** function has been used which implements the Tree-Augmented naive Bayes classifier. BC_1 represents classic Naive Bayes algorithm, whose structure (DAG) is fixed and must not be learned from the data. The estimation of the other parameters, for the other classifiers, are got using the *maximum likelihood estimation* (MLE) method. We used *gRain* package [44] to carry on the Bayesian inferences.

We make some comparisons among the ensemble we construct from the five Bayesian classifiers and other classifiers from state-of-the-art: Neural Network (NN), Support Vector Machine (SVM) and Random Forest (RF), which have been constructed, respectively, with the functions **mlNnet**, **mlSvm** and **mlRforest** of the *mlearning* package of R,[1] by using the default values in the first two (maximum number of iterations = 1000 for NN and radial kernel for SVM), and 5 trees to generate for RF.

In the literature on predicting the risk of mortality for hospital patients, it is common to use APACHE II, which is a severity of disease classification system that uses basic physiologic principles, to stratify acutely ill patients prognostically by risk of death. The standard approach (see for example [9]) is to compute the individual risk of death (probability of "die" for the variable *Result*) as

$$\frac{e^{\text{logit}}}{1 + e^{\text{logit}}},$$

where *logit* is obtained from the following equation:

$$\begin{aligned}\text{logit} = {} & -3.517 + 0.146 \times \text{APACHE II (the numeric value of } F_{21}) \\ & + 0.603 \text{ (only if post-emergency surgical, that is, } F_{18} = \text{`Urgent Surgical'}) \\ & + \text{coefficient } \beta \text{ (weight) of the diagnostic categories } F_4 \text{ to } F_{15} \text{ and } F_{19}.\end{aligned} \tag{2}$$

(coefficients β are fixed and have been recorded in Appendix D). When we follow this approach the data is just used for validation, not for training. With the intention of improving this classifier, we construct a locally recalibrated model based on APACHE II using the same training sets used for the base classifiers BC_1 to BC_5, NN, SVM, RF, and the ensembles EWA, EA, WMV and MV, on which the coefficients in Eq. (2) are learned on each training dataset, instead to be fixed. We name this model by LR.APACHEII, and it is built using *logistic regression*, implemented by the function **glm** of R (with argument *family* equal to "binomial"), from the same data used to construct the other classifiers, with regressors: F_4 to F_{15}, F_{18} and F_{19}. The individual risk of death (probability of "die" for the variable *Result*) with the model LR.APACHEII has been computed as

[1] Grosjean, Ph., Denis, K.; (2013) mlearning: *Machine learning* algorithms with unified interface and confusion matrices. R package version 1.0-0. https://CRAN.R-project.org/package=mlearning.

Table 7
Confusion matrices, with the corresponding Accuracy (Acc) and F-score (F), for the fourteen classifiers predicting output variable *Result*, for the first run and the first fold. Predictions are given by row, and observed classes by column, with the order of the classes, +"die", −"live". NaN means "not a number".

$BC_1: \begin{pmatrix} 22 & 28 \\ 4 & 196 \end{pmatrix}$ Acc=0.872, F=0.2895	$BC_2: \begin{pmatrix} 14 & 17 \\ 12 & 207 \end{pmatrix}$ Acc=0.844, F=0.491	$BC_3: \begin{pmatrix} 0 & 0 \\ 26 & 217 \end{pmatrix}$ Acc=0.893, F=NaN	$BC_4: \begin{pmatrix} 11 & 9 \\ 15 & 211 \end{pmatrix}$ Acc=0.902, F=0.478	$BC_5: \begin{pmatrix} 10 & 12 \\ 16 & 208 \end{pmatrix}$ Acc=0.886, F=0.417
$NN: \begin{pmatrix} 0 & 0 \\ 26 & 224 \end{pmatrix}$ Acc=0.896, F=NaN	$SVM: \begin{pmatrix} 7 & 5 \\ 19 & 219 \end{pmatrix}$ Acc=0.904, F=0.368	$RF: \begin{pmatrix} 8 & 13 \\ 18 & 211 \end{pmatrix}$ Acc=0.876, F=0.340		
$APACHEII: \begin{pmatrix} 0 & 0 \\ 19 & 200 \end{pmatrix}$ Acc=0.913, F=NaN	$LR.APACHEII: \begin{pmatrix} 2 & 1 \\ 17 & 199 \end{pmatrix}$ Acc=0.918, F=0.091			
$MV: \begin{pmatrix} 11 & 11 \\ 15 & 213 \end{pmatrix}$ Acc=0.896, F=0.458	$WMV: \begin{pmatrix} 11 & 10 \\ 15 & 214 \end{pmatrix}$ Acc=0.900, F=0.468	$EA: \begin{pmatrix} 15 & 11 \\ 11 & 213 \end{pmatrix}$ Acc=0.912, F=0.577	$EWA: \begin{pmatrix} 16 & 11 \\ 10 & 213 \end{pmatrix}$ Acc=0.916, F=0.604	

$$\frac{e^{\text{LR.logit}}}{1+e^{\text{LR.logit}}},$$

where LR.*logit* is obtained from the following equation, learned from data:

$$\begin{aligned} \text{LR.logit} = \; & \alpha_0 + \alpha_1 \times \text{APACHE II (the numeric value of } F_{21}) \\ & + \alpha_2 \times F_{18} + \alpha_3 \times F_{19} + \sum_{j=4}^{15} \alpha_j \times F_j. \end{aligned} \tag{3}$$

That is, the coefficients α of the diagnostic categories are learned from data. In our case, when we learn the model from the complete dataset, the corresponding coefficients α have been recorded in Table 34 in Appendix D, with the *p*-values for statistical significance. Referring to mortality, in Table 34 we can see what is the only protection factor, which is F_4 (in boldface), and what are the risk factors (the remaining in the table).

2.4. Validation and comparison with other classifiers

We choose to carry out the process of $k-$fold cross-validation with $k=10$ folds to validate our proposed hierarchical model. We use four different performance metrics to make comparisons between the EWA classifier and its single component base classifiers BC_1, …, BC_5, as well as with the other *pattern recognition* methods: NN, SVM and RF, and with the ensembles EA and that based on the fusion of labels outputs introduced in [37], MV and WMV. If a tie takes place with the latter, what happens when the evidence consisting of the patient's features has an estimated probability equal to zero with any of the five base classifiers, the tiebreaker rule will assign one of the categories at random, with equal probabilities. A further comparison is against the classifiers based on the APACHE II score following the traditional approach, the usual one and the enhancement we proposed, which is the locally recalibrated LR.APACHEII.

We randomize in order to reduce the possible bias due to the (random) choice of the folds in the validation process. Moreover, we repeat the process 20 times, using a different seed (randomly selected) in each case to carry out the partition of the database into the $k=10$ folds. The metrics used to make comparisons between the classifiers are:

–**AUPR:** as we have already commented, this is considered a good measure when the database is unbalanced with respect to the class variable, as is our case.

–**F-score:** our goal has been to enhance the prediction of the minority class (identified with the "positive" class). For that, our interest is focused on the improvement of *sensitivity* (*recall*), which together with *precision* are the two measures that make up the F-score, defined as their harmonic mean.

–**AUC (Area Under the ROC[2] Curve):** very popular in the medical literature. An advantage of incorporating it as a metric in the validation process is that the results obtained in our study may be compared with those of others made with different populations and methodologies.

We must highlight the imbalance of the class distribution in the case of the output variable *Result*, with minority class "die" (14.7%), and also in the case of output variable *Destination*, with minority class "Major Complexity Hospital" representing a 8.9% (of the cases with known destination and different from "Morgue"). In the case of the output variable *Cause*, the minority class "Complications" represents 21.6% of the cases with known cause of death, so the imbalance is not so extreme.

Although it is the most common of the metrics, we do not include the *Accuracy* in this study because it is not very suitable in cases of imbalance by the *accuracy paradox*. In Table 7 below we report as illustrative example the confusion matrices obtained in the validation procedure for each the fourteen classifiers predicting output variable *Result*, for the first run and the first fold, jointly with the corresponding Accuracy (Acc) and F-score (F) values. In the matrices, predicted classes are given by row, while observed by column, in order: +"die", −"live". Note that the number of observed cases for APACHEII and LR.APACHEII models is 219 while for the rest is 250; the reason is that the first cannot provide any prediction if F_{18}, F_{19} or F_{21} are missing. Although it is only an example and the matrices are subject to variability, from them we can get an idea of what is happening with the different models: APACHEII always predicts "live" (BC_3 and NN only do it sometimes), having exactly the Accuracy given by the proportion of the majority class in the validation set, and F-score cannot be computed. The rest of classifiers sacrifice the correct prediction of all the majority class in order to be able to correctly predict some patients of the minority class, that is, patients that died, which is what we are interested in from a clinical point of view. However, the Accuracy of the ensembles EA and EWA is comparable to that of APACHEII and LR.APACHEII, having a higher F-score value. This idea is confirmed with the statistical comparison among them explained in Section 3.

[2] The ROC, Receiver Operating Characteristic curve serves to illustrate the capacity of diagnostic of a binary classifier as the discrimination threshold varies; it plots the *sensitivity* (or True Positive Rate) against the False Positive Rate (1-*specificity*).

Table 8
Average over the runs of the averages ($\overline{x}$) and the standard deviations (s) over the folds, for the metrics AUPR, F-score and AUC, with the different classifiers. Output variable *Result*. In boldface, the top five for each metric.

Result	AUPR		F-score		AUC	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
BC_1	**0.52058 (5)**	0.08914	**0.54445 (1)**	0.06056	**0.87230 (3)**	0.02858
BC_2	0.50671	0.08514	**0.52763 (3)**	0.06642	0.85987	0.03192
BC_3	0.35161	0.13386	0.06805	0.02496	0.82825	0.04087
BC_4	0.48450	0.07944	0.49974	0.06621	0.83958	0.03605
BC_5	0.46424	0.07927	0.47569	0.06794	0.83277	0.03456
NN	0.27294	0.23893	0.43670	0.07689	0.70228	0.18079
SVM	0.43432	0.08309	0.32713	0.08628	0.79698	0.04014
RF	0.37071	0.08110	0.37567	0.08028	0.76864	0.04096
APACHEII	0.37899	0.09393			0.77518	0.04837
LR.APACHEII	0.42621	0.09342	0.30706	0.09075	0.83154	0.03744
MV	**0.52467 (3)**	0.08276	0.50274	0.06744	**0.86440 (4)**	0.03027
WMV	**0.52317 (4)**	0.08316	**0.51137 (5)**	0.06791	**0.86377 (5)**	0.03098
EA	**0.53829 (2)**	0.08510	**0.52354 (4)**	0.06666	**0.87913 (2)**	0.02538
EWA	**0.54131 (1)**	0.08423	**0.53270 (2)**	0.06766	**0.88026 (1)**	0.02522

3. Results

3.1. For the variable Result

The boxplots in Fig. 6 (Appendix A) correspond to the values of AUPR, F-score and AUC obtained by using k-fold cross validation with $k = 10$, for prediction of the output variable *Result* with *positive* class "die", for the first run and the fourteen classifiers considered in the study (including the two based on APACHE II for predicting mortality at the ICU level, which are the standard one and the locally recalibrated LR. APACHEII).

We record the average over the 20 runs of the averages and the standard deviations, $\overline{x}$ and s, respectively, over the folds, for AUPR, F-score and AUC when considering the output variable *Result*, in Table 8. The blank cells indicate that the F-score could not be calculated by the arrangement of the zeros in the confusion matrices generated by the APACHEII model.

From the experiment, we can see that there is a clear advantage for the ensembles, especially EWA and EA, over the rest of the classifiers that have been considered, using AUPR and AUC as performance measures while for the F-score, the best classifiers are BC_1, EWA, BC_2 and EA. That is why we will focus on the comparison between the ensembles EWA, EA, WMV and MV, to each other, in addition to in their comparison with the rest. For each of the metrics, below we detail some of the results.

AUPR: Table 17 (Appendix B) reports for each run if there is a statistically significant (p-value <0.1) improvement of either EWA or EA, with respect to WMV and/or MV. Here "2" means that there is an improvement over WMV and MV, "1" means that there is only an improvement over one of them, and "0" that there is none for either. In no case are WMV or MV better than EWA or EA. p-values[3] are reported in Table 18 and have been adjusted for multiple comparisons between the four ensembles by using the method of Holm-Bonferroni, with the pairwise Wilcoxon signed-rank test [45] to compare matched pairs of samples corresponding to the same run. This statistical test is used as an alternative to the Student's t-test when the population cannot be assumed to be normally distributed (according to the Shapiro-Wilk test [46], which has been previously performed).

From these tables se see that EWA and EA outperform WMV and MV, and that in 5 runs, there are significant differences among EWA and EA and, in all the cases, EWA shows to be better. This is confirmed in Table 19, where we observe that EWA is significantly better than EA in 8 runs, when we compare only the two and, therefore, the p-values have not been adjusted, and in all the cases, EWA shows to be better. What significance does this fact have? We compute the p-value for the exact Binomial test in order to compare the proportions of cases in which EWA outperforms EA and vice versa, instead of use McNemar test, because the sample is small. The one-sided p-value for the exact Binomial test is $P(B(n = 5, p = 0.5) = 5) = 0.5^5 = \mathbf{0.03125}$* when we compare EWA and EA but adjust for the comparison of the four ensembles, which decreases to $P(B(n = 8,\ p = 0.5) = 8) = 0.5^8 = \mathbf{0.00391}$** if we consider the non-adjust corresponding to comparison of EWA against EA alone. In both cases there is a statistically significant evidence in favour or EWA as opposed to EA for prediction of variable *Result*, with AUPR as performance measure.

As regards APACHEII and LR.APACHEII, both are clearly worse than any of the ensembles, and we observe significant differences among them, in favour of the latter. Indeed, for 18 runs (see details in Table 20 in Appendix B) there are differences between the medians for the AUPR metric, and in all cases LR.APACHEII turns out to be better than the standard based on APACHE II, with a one-sided p-value for the exact Binomial test: $0.5^{18} = \boldsymbol{3}.81470 \times 10^{-6}$***.

F-score: We repeat the procedure with the F-score and obtain Tables 21–23 in Appendix B. We observe that EWA and EA outperform WMV and MV (and also that WMV behaves better than MV), and that EWA is better than EA. From Table 23 we have that in 14 runs there are significant differences among EWA and EA, when comparing the only two, and in all the cases, EWA is better, which has the following one-sided p-value for the exact Binomial test: $0.5^{14} = \boldsymbol{6}.10352 \times 10^{-5}$***, which gives a clear statistical significance in favour of EWA against EA.

If we compare the best of the ensembles, EWA, against the base classifiers (with the adjusted p-values for multiple comparisons) we see that the median of the F-score of both EWA and BC_1 is significantly greater than that of BC_4 and BC_5, and that the median of BC_2 is significantly greater than that of BC_5. We can no obtain statistical significance with respect to BC_3, which has the lowest F-score values due to the large number of missing values. They are not observed either significant differences between EWA, BC_1 and BC_2 as far as the F-score is concerned, if we compare the three with each other as a single block, and even if we compare independently in pairs.

The locally recalibrated LR.APACHEII is clearly worse than the four ensembles when using the F-score as performance metric. Note that it is not possible to calculate the F-score for the standard based on APACHE II since in all cases the prediction with this classifier for the output variable *Result* was the majority class "live", resulting in degenerate confusion

[3] As usual, throughout the paper · denotes significance at 10%, superscript * denotes statistical significance at 5%, ** at 1% and *** at 1‰, for all the p-values.

Table 9
Average over the runs of the averages ($\bar{x}$) and the standard deviations (s) over the folds, for the metrics AUPR, F-score and AUC, with the different classifiers. Output variable *Destination* In boldface, the top five for each metric.

Destination	AUPR		F-score		AUC	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
BC_1	0.26687	0.10371	0.28345	0.11281	**0.75532 (4)**	0.07043
BC_2	0.27571	0.09357	0.26376	0.10863	**0.74604 (5)**	0.07077
BC_3	**0.36120 (1)**	0.11843	**0.33922 (1)**	0.12737	**0.75596 (3)**	0.06692
BC_4	0.26142	0.09548	**0.31516 (2)**	0.11396	0.70311	0.07645
BC_5	0.24706	0.09042	0.26310	0.11213	0.70424	0.07592
NN	0.16834	0.11205			0.67487	0.10562
SVM	0.22107	0.08253	0.13317	0.04624	0.66455	0.07480
RF	0.21222	0.07203	0.21634	0.09616	0.65715	0.066328
MV	**0.29333 (4)**	0.10329	**0.30076 (4)**	0.12116	0.73914	0.07266
WMV	**0.29598 (3)**	0.09814	**0.30463 (3)**	0.12205	0.73493	0.07158
EA	**0.28726 (5)**	0.10200	0.28635	0.11838	**0.76802 (2)**	0.057239
EWA	**0.29766 (2)**	0.10470	**0.29444 (5)**	0.11912	**0.77201 (1)**	0.05715

matrices with a row equal to zero.

AUC: Comparing the fourteen classifiers all at one, we see that the ensembles jointly with BC_1 and BC_2 are the best, the rest being far below. Then, first we compare the ensembles between them. In Table 24 (Appendix B), which is analogous to Table 17, we report for each run if there is a statistically significant (p-value <0.1) improvement of either EWA or EA, with respect to WMV and/or MV, and the corresponding p-values are in Table 26 (note that they have been adjusted for multiple comparisons between the four ensembles). We see clearly that EWA and EA outperform WMV and MV, and that EWA does with respect to EA in 6 runs. Previously, in Table 25 we record the results of the comparison between only EWA and EA, and in 12 runs EWA shows to be better (one-sided p-value for the exact Binomial test: $0.5^{12} = \boldsymbol{0}.00024$***). Consequently, EWA is the best of the ensembles.

It only remains for us to compare it with BC_1 and BC_2. In comparing the three at one, with the corresponding adjustment of the p-values, we observe that BC_1 is significantly better than BC_2 in 8 runs, and both are worst than EWA (in 5 and 18 of the runs, respectively, with one-sided p-values for the exact Binomial test: $0.5^{5} = \boldsymbol{0}.03125$* and $0.5^{18} = \boldsymbol{3}.8 \times 10^{-6}$***). Definitely, then, EWA is the best of the classifiers for output variable *Result* attending to AUC metric.

As regards comparison between APACHEII and LR.APACHEII, the latter turns out to be better in all the runs, with a one-sided p-value for the exact Binomial test: $0.5^{20} = \boldsymbol{9}.5367 \times 10^{-7}$***.

3.2. For the variable Destination

Fig. 7 in Appendix A shows the boxplots for AUPR, F-score and AUC for first run and the twelve classifiers, for prediction of the output variable *Destination* with *positive* class "Major Complexity Hospital". Analogous results to that of Table 8 are in Table 9. Below we specify a little more the results by metrics.

AUPR: Regarding the comparison between ensembles, Table 28 in Appendix C shows the adjusted p-values for multiple comparisons between the four ensembles MV, WMV, EA and EWA. We can see that EWA clearly outperforms EA, and WMV also outperforms EA in two runs. Table 29 refers to the comparison between EWA and EA alone (non-adjusted p-values), showing again that the former beats the last one in 19 of the runs. The corresponding one-sided p-value for the exact Binomial test in favour of EWA is $0.5^{19} = \boldsymbol{1}.90735 \times 10^{-6}$***.

However, the classifier that performs the best in general is not one of the ensembles but BC_3. See Table 30 in Appendix C where the comparison of BC_3 with two different groups of classifiers is shown: the base classifiers on the one hand, and the ensembles on the other. It is observed that BC_3 surpasses them all.

Table 10
Average over the runs of the averages ($\bar{x}$) and the standard deviations (s) over the folds, for the metrics AUPR, F-score and AUC with the different classifiers. Output variable *Cause*. In boldface, the top five for each metric.

Cause	AUPR		F-score		AUC	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
BC_1	**0.28843 (2)**	0.13782	0.35156	0.13504	0.66947	0.13908
BC_2	**0.26481 (5)**	0.14900	0.40100	0.15465	**0.68449 (3)**	0.15798
BC_4	0.23687	0.14077	0.41818	0.14623	0.64460	0.21162
BC_5	0.19745	0.14208	**0.49238 (1)**	0.17201	0.53813	0.21570
NN	0.19639	0.14891			0.55821	0.10866
SVM	**0.30422 (1)**	0.12329			0.62733	0.11411
RF	0.18800	0.09173	0.42469	0.11681	0.52595	0.12109
MV	**0.27028 (4)**	0.15662	**0.48086 (2)**	0.16753	**0.70198 (1)**	0.17819
WMV	**0.27764 (3)**	0.16227	**0.46805 (3)**	0.15084	**0.69832 (2)**	0.17886
EA	0.26473	0.15841	**0.43952 (5)**	0.13424	**0.68046 (4)**	0.19690
EWA	0.26479	0.15690	**0.44878 (4)**	0.13594	**0.67563 (5)**	0.19324

F-score: With regard to the F-score metric, there is no significant differences considering all the classifiers at once (although for some runs RF shows to have a lower F-score than classifiers BC_3, BC_4 or the ensembles), nor considering the four ensembles together, but there are if we consider EWA and EA alone, showing that EWA is the best of both. The corresponding p-values are in Table 31, showing that in 7 runs, EWA outperforms EA; the one-sided p-value for the exact Binomial test in favor of EWA is $0.5^{7} = \boldsymbol{0}.00781$**. Regarding BC_3, there are no significant differences between this classifier and the ensembles, when we compare the five together (using adjusted p-values), although there are if we compare in pairs (BC_3 vs. each ensemble).

AUC: If we compare the twelve classifiers all at one, we see that the ensembles EWA and EA are significantly better than NN, SVM and RF, slightly better than BC_4 and BC_5, and no differences have been observed with BC_1, BC_2 and BC_3. For that, we decide to compare as a block these last with EWA and EA, since the adjustment of the p-values when making a large number of multiple comparisons can mask differences that are

really significant. There is only a slight evidence in favour of EWA, since it proves to have a significantly higher AUC median than BC_2 and EA in 3 of the runs, while none of the other classifiers beat it in any run, and if we compare in pairs (without adjusting the *p*-values), EWA outperforms each of EA, BC_1, BC_2 and BC_3 (see Table 32) so it turns out to be the favorite, although relative to BC_3, it is only slightly higher.

3.3. For the variable Cause

The boxplots for AUPR, F-score and AUC for first run and the twelve classifiers, for prediction of the output variable *Cause* with *positive* class "Complications", are in Fig. 8 (Appendix A). The average over the 20 runs of the averages and the standard deviations over the folds, for the output variable *Cause*, are recorded in Table 10 below.

There are few significant differences between the classifiers for any of the metrics (AUPR, F-score and AUC), so the results do not seem conclusive, and we will have to wait for more data. The problem with the output variable *Cause* is that, due to the scarcity of cases in which the patient died in ICU (367 patients in our cohort), there are a large number of zeros in the confusion matrices, and consequently, a large number of missing values in the metrics. For example, if we compare AUPR for the classifiers SVM, EWA and EA with the pairwise Wilcoxon signed-rank test and making Holm-Bonferroni adjustments for multiple comparisons, we cannot see significant differences. We have to make comparisons in pairs to find something. Specifically, in 4 of the runs the median of SVM is statistically significantly greater than that of EWA (one-sided *p*-value for the exact Binomial test: $0.5^4 = \boldsymbol{0}.0625\cdot$), while this number increases to 6 for the comparison with EA instead of EWA (*p*-value $0.5^6 = \boldsymbol{0}.015625^*$).

4. Centrality, odds ratio and feature strength

Centrality and betweenness: Directed Acyclic Graphs (DAGs) from Figs. 9 and 10 represent the relationships of conditional independence entailed by the Bayesian networks BC_2, BC_3, BC_4 and BC_5, which are helpful to interpret the EWA ensemble, defined as an ensemble of them, jointly with BC_1, which is a Naive Bayes. We can establish which features play the main role in the model by using centrality and/or betweenness measures borrowed from the Network Analysis area applied to the DAGs, as it was done in [47]. In Graph Theory and Network Analysis, indicators of centrality identify the most important (influential) nodes within a graph, where "importance" is conceived as involvement in the cohesiveness of the network. For each feature we compute four different of these indicators (see [48]), which are shown in Tables 35 and 36 (Appendix E), normalized in order to sum up 100:

(a) *Freeman's degree of centrality*, which counts paths which pass through each node, that is, directed arcs which arrive at or depart from it.
(b) *Basic standard betweenness measure*, which quantifies the number of times a node acts as a "bridge" along the shortest path between two other nodes (which we will call "geodesic" from now on). Nodes that have a high probability to occur on a randomly chosen geodesic between two randomly chosen nodes, have a high betweenness. Fixed a node v, this measure is defined by $\sum_{i,j,i\neq j,i\neq v,j\neq v}(g_{ivj}/g_{ij})$ (with the convention $0/0=0$), where g_{ij} is the number of geodesics from i to j in the graph, and g_{ivj} is the number of geodesics in the subset of those that pass through v.
(c) *Borgatti's proximal source betweenness* is a variant of basic standard betweenness to accumulate only for the last intermediating vertex in each incoming geodesic; this expresses the notion that, by serving as the "proximal source" for the target, this particular intermediary node will in some settings have greater influence than the rest. Fixed a node v, this measure is defined by $\sum_{i,j,i\neq j,j\neq v,i\rightarrow v}(g_{ivj}/g_{ij})$
(d) *Borgatti's proximal target betweenness* is the counterpart to proximal source betweenness that allows betweenness to accumulate only for the first intermediating vertex in each outgoing geodesic; this expresses the notion that, by serving as the "proximal target" for the source, this particular intermediary node will in some settings have greater influence or control than others. Fixed a node v, this measure is defined by $\sum_{i,j,i\neq j,i\neq v,j\leftarrow v}(g_{ivj}/g_{ij})$

Table 11
The most influential features attending to centrality and betweenness.

Demographic characteristics	F_2: Age
Main cause of admission	F_4: ACS F_5: RF F_6: Shock
Admission	F_{18}: Generic Syndrome F_{19}: Sepsis
Severity (on first 24 h of admission)	F_{20}: ICU Workload F_{21}: APACHE II

Table 12
Example of characteristics of a critically ill patient.

F_3: Charlson	F_{17}: Origin	F_{18}: Generic syndr.	F_{19}: Sepsis	F_{20}: ICU Workload	F_{21}: APACHE II
2	Emergency Room	Medical	Yes	M. unstable coma/shock	5–9

Features F_9 (Hepatic F), F_{13} (CT), F_{14} (OT) and F_{15} (Intoxication), all of them corresponding to the category of "Main cause of admission", do not appear in Tables 35 and 36 because their value of the three betweenness variants is 0 for BC_2, BC_3, BC_4 and BC_5, and at the same time, their Freeman's centrality value is very small. They are, therefore, the least important for the cohesiveness of the network, which is logical since none of them exceeds 1.5% of prevalence in the cohort. At the other extreme, there are the most important features in this regard, with higher values of centrality and betweenness (see Table 11): they are the most influential for the predictive model but in relation with others. Note that those that are in the "Main cause of admission" category, are the three most prevalent.

Features in Table 11 act as gateways, and the arcs that connect them as bridges, through which information flows from one cluster of variables in the model to another. We can see from the DAG of BC_3 in Fig. 9 (we consider this DAG because it is built without forcing any directed arcs, see Table 3) that

–The link between F_4 and F_{18} is a bridge, and while F_4 is a gateway to F_{21} and to the demographic characteristics, F_{18} is to F_{20}, to the main features of the "Main cause of admission" category, and to that of "Admission".
–The information between the two clusters mentioned in the above item, "Demographic Characteristics" and "Comorbidities" on one hand, and "Admission" on the other, also flows through the concatenation of two bridges: between F_{20} and F_{21}, the features of "Severity

Table 13
Example of Table 12: probabilities of "die" and odds ratio in favour of "die", for each of the possible "Main cause of admission". In boldface those probabilities >0.5, which carry a prediction of "die" for the patient.

F_1: Sex	Male			Female		
F_2: Age	75–84	>84	OR	75–84	>84	OR
F_4	0.07878	0.09715	1.25837	0.08084	0.10693	1.36138
F_5	0.19627	0.30552	1.80145	0.22859	0.36471	1.93737
F_6	0.20421	0.31516	1.79328	0.23714	0.37257	1.91022
F_7	0.20070	0.31982	1.87257	0.23325	0.37825	1.99987
F_8	0.46216	**0.55110**	1.42871	**0.50267**	**0.57898**	1.36059
F_9	0.16010	0.16010	1.00000	0.16010	0.16010	1.00000
F_{10}	0.49500	**0.66774**	2.05030	**0.53169**	**0.72091**	2.27516
F_{11}	0.20996	0.26928	1.38660	0.23044	0.31191	1.51378
F_{12}	0.07956	0.13082	1.74141	0.09336	0.16242	1.88317
F_{13}	0.30486	0.38341	1.41787	0.35674	0.44058	1.42009
F_{14}	0.07100	0.07003	0.98518	0.07433	0.07305	0.98137
F_{15}	0.10251	0.20942	2.31921	0.11517	0.22107	2.18048
F_{16}	0.08263	0.10638	1.32156	0.08962	0.12278	1.42184

(on first 24 h of admission)", and between F_{21} and F_2. The latter is very natural since the APACHE II score is calculated based on age.
–Within the cluster of features of "Admission", there is a bridge between F_5 and F_6, connecting sub-clusters; F_5 is a gateway to F_{17}: Origin and F_{19}: Sepsis, while F_6 is to F_{10}: CRA and F_{16}: Other syndromes.

Odds Ratio: Besides, we can use EWA, which has proven to be the best of those we have considered, to evaluate the effect of the features in the evaluation of the risk of death. For example, for each of the "Main cause of admission" we can compute the Odds Ratio (OR) in favour of "die" when the feature is present compared to when it is not, being the other absents. An odds ratio (OR) is a measure of association between a feature and the outcome (variable *Result*, in this case), which represents the odds in favour of "die" given a particular value of a feature, compared to the odds in favour of "die" occurring given another value. For that, we fix the other features.

Just as an example of this, consider a critically ill patient with the characteristics in Table 12, in the year 2018.

In Table 13 we record the odds ratio, disaggregated by sex, in favour of the event "die", for a critically patient whose characteristics are given in Table 12, according to what of the "Main cause of admission" has been reported for the patient (from F_4 to F_{16}). The odds ratio is defined as the ratio of the odds of event "die" occurring in the group of age >84 to the odds of it occurring in the group of age 75–84. Let continue with the example of the patient whose characteristics are given in Table 12: a male between 75 and 84 years old and with renal failure ($F_8 = 1$). Thus, the risk of death (probability of "die") is 0.46216. This probability increases up to 0.55110 if the age increases to be > 84. Therefore, the Odds Ratio in favour of "die" is:

$$\mathrm{OR}_{>84/75-84} = \frac{0.55110/(1-0.55110)}{0.46216/(1-0.46216)} = 1.42871$$

With respect to the risk of death, we observe the following, which is consistent with what is observed in Figs. 2 and 3 in Appendix A:

–it is greater for women than for men, for both intervals of age 75–84 and >84 and for all of the "Main cause of admission" features except for F_9, in which case no variation in risk is observed (F_9, Hepatic Failure, is one of the features less important from the perspective of centrality, we have seen).

Table 14
SD, the correction term δ and CSD for the 21 features.

Feature	SD	δ	CSD($=$ SD $\times\ \delta$)
F_1	0.03432	1/2	0.01716
F_2	0.09623	1/2	0.04812
F_3	0.14104	1/2	0.07052
F_4	0.17367	1/2	0.08684
F_5	0.03301	1/2	0.01651
F_6	0.08698	1/2	0.04349
F_7	0.02511	1/2	0.01256
F_8	0.00865	1/2	0.00433
F_9	0.13364	1/2	0.06682
F_{10}	0.49694	1	0.49694
F_{11}	0.15203	1/2	0.07602
F_{12}	0.09849	1/2	0.04925
F_{13}	0.15729	1/2	0.07865
F_{14}	0.12650	1/2	0.06325
F_{15}	0.07601	1/2	0.03801
F_{16}	0.12105	1/2	0.06053
F_{17}	0.35811	1/2	0.17906
F_{18}	0.17512	1/2	0.08756
F_{19}	0.11333	1/2	0.05667
F_{20}	0.43213	1/2	0.21607
F_{21}	0.63546	1	0.63546

–within each sex, it is greater for older people, except when Other Trauma (F_{14}) is present, case in which the variation is very small (the same thing happens than with F_9).

We also see that for patients having the characteristics recorded in Table 12 and having cardio respiratory arrest (F_{10}: CRA) or intoxication (F_{15}: Intoxication), both for men and women, the increase in age is an important risk factor (OR greater than 2 in Table 13).

On the other hand, we can study which of the "Main cause of admission" are risk factors for a male who is more than 85 years old, and with the features in Table 12, for example, and consider the question: "What is the Odds Ratio between F_{10}: CRA and F_5: RF in favor of *die*?", which is answered by computing the ratio between the odds in favour of "die" when $F_{10} = 1$ and when $F_5 = 1$, which is:

$$\mathrm{OR}_{F_{10}/F_5} = \frac{0.66774/(1-0.66774)}{0.30552/(1-0.30552)} = 4.56836$$

(see Table 13) that is, the odds in favour of "die" if $F_{10} = 1$ (Cardio Respiratory Arrest) is approximately 4.6 times greater than if $F_5 = 1$ (Respiratory Failure), for a patient with the mentioned characteristics, on which this result may depend, obviously.

Feature strength: Finally, we compute a measure of the feature strength to predict the output *Result*. For that, we follow [49] and introduce a measure based on the conditional probability tables of *Result* with respect to each feature, obtained with EWA (see Appendix F). This measure uses the Kolmogorov-Smirnov statistical distance and a correction parameter. Indeed, we first introduce a strength measure for each feature, say F, named *Strength Distance* (SD), in this way:

$$\mathrm{SD}(\mathrm{F}) = \max_{a,b\in\mathcal{F}} d^{\mathrm{F}}_{a,b}$$

where $\mathcal{F}$ is the set of the possible outcomes of variable F, and $d^{\mathrm{F}}_{a,b}$ denotes the Kolmogorov–Smirnov statistical distance between the *a posteriori* conditional probability distributions of *Result* given the evidence $\mathrm{F} = a$, and given the evidence $\mathrm{F} = b$. The values of SD have been recorded in Table 14 below. To take into account if different instantiations of a feature produce different predictions for *Result*, we introduce the correction term $\delta(\mathrm{F}) = \gamma(\mathrm{F})/2 \in (0,\ 1]$, where $\gamma(\mathrm{F})$ is the number of different predictions obtained from the classifier for *Result* given the

Table 15
Best classifier(s) for the output variables *Result* and *Destination* and performance metrics, according to our experimental evaluation. If the classifiers are not in boldface, it means that only slightly exceeds its competitors. In each scenario, it is indicated by Yes/No if LR.APACHEII is significantly better than APACHEII, when the comparison makes sense.

		Performance metric		
		AUPR	F-score	AUC
Output	Result	**EWA**	BC_1,EWA,BC_2	**EWA**
		Yes		Yes
	Destination	**BC_3**	**BC_3**	EWA

evidences of the form $F = a$, with a varying in $\mathcal{F}$. Then, $\delta(F)$ is the proportion of different predictions actually obtained by the classifier for *Result* among the possible we could obtain from an evidence on F, which is 2, and we use it to correct strength measure SD by introducing the *Corrected Strength Distance* (CSD) by $CSD(F) = SD(F) \times \delta(F)$. Note that $CSD(F) \geq 0$, and that $CSD(F) = 0$ if and only if F and *Result* are independent variables. In Table 14 we have recorded for each feature the correction term δ and the feature strength measure CSD as well.

Attending to CSD as feature strength measure, we can rank the features as follows, from stronger to weaker, attending to their capacity to modify prediction of the output *Result*:

$$F_{21}, F_{10}, F_{20}, F_{17}, F_{18}, F_4, F_{13}, F_{11}, F_3, F_9,$$
$$F_{14}, F_{16}, F_{19}, F_{12}, F_2, F_6, F_{15}, F_1, F_5, F_7, F_8.$$

5. Conclusions

It is unlikely that intelligent software will replace the clinician in medical diagnosis and prognosis for patients care. *Machine learning* expert systems are more likely to act as intelligent agents for specialized, complicated problems, and are intended to enhance the performance of the expert physician, given place to smart Intensive Care Units in the future. Our primary goal in this work has been to demonstrate the feasibility and benefits of routinely collecting information from critically ill patients admitted to ICU. The development of automatic tools to assist in clinical decision-making remains a challenge, although steps have already been taken in this direction. Our research is directed towards the construction of a *machine learning* hierarchical classifier to predict the risk of death in the ICU, as well as destination, for those who survive their stay in the ICU, or the cause of death for the rest, from multiple data streams ("Demographic Characteristics", "Comorbidities", "Admission" and "Severity (on first 24 h of admission)").

In a first step, an ensemble of Bayesian classifiers (EWA) is developed to predict the risk of death (output variable *Result*), while another is used to predict the destination of the patient at ICU discharge if the predicted value for the variable *Result* is *live*, or his/her cause of death if the prediction is *die*. EWA is constructed as an ensemble of five different base Bayesian networks, with the weighted average rule with appropriate weights, which are obtained from the estimations of the AUPR values of the base classifiers to give more power to more "competent" base classifiers in the average criterion. When dealing with highly unbalanced and sparse datasets, AUPR, F-score and AUC are preferred as representation of performance assessment in the binary classification to the most commonly used measure, the *Accuracy*.

We compare the performance of EWA with that of the base Bayesian classifiers from which it has been constructed, and with some state-of-the-art *machine learning* methodologies (Neural Networks, Support Vector Machine and Random Forest), as well as with the ensembles of the same base classifiers obtained using the average rule without weights, the majority vote and the weighted majority vote criteria, finding that EWA has best overall performance. We also see as expected, that EWA improves the models based on scales, such as the traditional approach based on the APACHE II score, which suffers from not incorporating elements that clinical practice reveals to be of great value, such as the origin of the patient (collected in our model in variable F_{17}, that has been ranked fourth in importance, according to CSD as measure of the feature strength), since depending on the origin, the patient may have a very different evolution in the ICU, presenting fragility to a greater or lesser degree. EWA even outperforms a local recalibration of this model obtained from the dataset, LR.APACHEII, both with the AUPR and AUC metrics, and between them, LR.APACHEII behaves better from a predictive point of view than APACHEII. When we consider the F-score metric, what happens is that, on the one hand, all the classifiers show better than LR.APACHEII while, on the other, it is not possible to calculate the F-score for the standard based on the APACHE II (*accuracy paradox*).

Table 16
Top five features, with the category that maximizes risk of death for each, and the associated risk.

Feature	Category that maximizes risk	Risk
F_{21}: APACHE II	>34	64%
F_{10}: CRA (Cardio Respiratory Arrest)	Yes	62%
F_{20}: ICU Workload	Medical unstable with coma or shock	44%
F_{17}: Origin	Extra Hospital Emergency	46%
F_{18}: Generic Syndrome	Medical	19.5%

Table 15 shows the best classifier for each output variable and performance metric. Note that *Cause* does not appear since for this output variable, there are no significant differences among the classifiers.

The conclusion from the statistical point of view of LR.APACHEII's superiority compared to APACHEII is that the logistic regression model based on the score APACHE II is a better predictor when estimates the parameters from the current database, which seems quite logical, since in this way the model better reflects the characteristics of the patients for whom the mortality risk prediction is intended and, in particular, it catchs the changes and improvements in medical practice. Indeed, APACHE II score was described and validated by means of a logistic regression model based on the management and results of critically ill patients in 1985, and since then, some obvious advances, in preventive and primary medicine or in the control of chronic diseases, have changed the relevance of age and comorbidities in the prognosis of critically ill patients, and in our century there has been a dramatic reduction in mortality due to sepsis, coronary syndromes and trauma. Therefore, healthcare professionals must take this into account when faced with the need to use scoring systems in their daily practice.

We also delve into interpretability of the EWA ensemble, both from the DAGs of the base classifier from which EWA has been constructed, using centrality and betweenness measures, and from the conditional probability tables of the outcome *Result* conditioned to any of the features, which allow us to rank the features attending to a measure of their strength. While this last approach discovers which features are important for prediction, features that are important but in relation with others, that is, the most "influential", stand out using centrality and betweenness, the rest of features being irrelevant for prognosis purpose.

The top five features are, in order of strength, in Table 16 below, with the categories that maximize the mortality risk, obtained from tables in Appendix F. Of these, F_{18}, F_{20} and F_{21} are also influential from the point of view of centrality and betweenness, so they appear as clearly

highlighted as basic characteristics to take into account to predict the risk of death in patients admitted to a hospital ICU. Note that a feature can be influential from the point of view of centrality and betweenness but have little predictive importance; for example, F_2: Sex has a relatively low value of CSD (meaning that the risk of death is similar for men and women, but it is influential since acts as connecting node between the features of the "Demographic Characteristics" and "Comorbidities" categories, which also have weak importance, and the rest of features.

Top five features are followed by F_4: ACS (Acute Coronary Syndrome), which is the only one characteristic of the "Main cause of admission" category that acts as factor of protection (its presence reduces the risk of death). From Table 38 we compute the OR in favour of "die" corresponding to the presence of ACS ($F_4 = 1$) with respect to its absence ($F_4 = 0$), regardless of the other features:

$$\mathrm{OR}_{F_4=1/F_4=0} = \frac{0.00390/(1-0.00390)}{0.11757/(1-0.11757)} = 0.02939$$

That is, the odds in favour of "die" divides by approximately 34 when ACS is present with respect to when it is absent, in general, without taking into account the other characteristics of the patient. This fact may seem counterintuitive, but it must be taken into account that clinical practice indicates that among the patients admitted to the ICU of a hospital, those who do so with "Generic Syndrome" F_{18} = "Coronary", which is clearly associated with ACS (F_4 = "yes"), are the ones with the best prognosis. Keep in mind, that if it is not due to that reason, the admission will be for another with a worse prognosis. For example, if they present a respiratory failure RF (F_5 = "yes"), which is associated with "Generic Syndrome" F_{18} = "Medical", their prognosis worsens (increases the risk of death), what is consistent with the information in Table 16. Note that 95.2% of the patients with F_{18} = "Coronary" present ACS, while only 0.7% of them present RF; from the patients with F_{18} = "Medical", 47.8% present RF but only 2.8% ACS.

Among the top five features ranked by strength, there is only one of the "Main cause of admission" category which seems to be the most important risk factor, F_{10}: CRA. From Table 40 we obtain similarly that the OR in favour of "die" corresponding to the presence of CRA ($F_{10} = 1$) with respect to its absence ($F_{10} = 0$) is

$$\mathrm{OR}_{F_{10}=1/F_{10}=0} = \frac{0.62298/(1-0.62298)}{0.12604/(1-0.12604)} = 11.45758$$

That is, the odds in favour of "die" multiplies by approx. 11.5 when CRA is present with respect to when it is absent. For a specific patient, based on its known characteristics, these result can be adjusted, as we have done in Section 4 in some cases to evaluate the effect of the features in the evaluation of the risk of death by means of the OR.

It would be very interesting to extrapolate our model to a database with a case mix of different ICUs, which would make it possible to compare the performance of different units; to do this, the characteristics of a typical patient would be introduced into the model and the risk of death would be predicted with each ICU. This tool would also make it possible to carry out a longitudinal study and analyze the improvement over time of the healthcare processes of a specific ICU, as well as adapting the model to the different types of ICU, from the trauma center to the thematic respiratory or cardiovascular ICU, if we learn it from data collected in these more specific scenarios.

To the extent that it can help physicians in undertaking patient-tailored therapeutic decisions, and to the health authorities to manage more optimally the available resources, the local data-driven *machine learning* methodology introduced in this work for estimating the risk of death and predicting the destination at ICU discharge or the cause of death, using an ensemble of Bayesian classifiers, seems to be a useful and promising tool with important clinical applicability.

Funding

R. Delgado and J.D. Núñez-González are supported by Ministerio de Ciencia, Innovación y Universidades, Gobierno de España, project ref. PGC2018-097848-B-I0.

R. Delgado, J.D. Núñez-González, Juan Carlos Yébenes and Angel Lavado are partially supported by TV3 Fundació Marató (Sepsis Training, Audit and Feedback (STAF) Project; Codi Projecte 201836).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

We want to acknowledge received contributions from reviewers during the review process and, especially Thanks to Editor-in-Chief Carlo Combi for the chance of having the article in consideration.

Appendix A. Plots

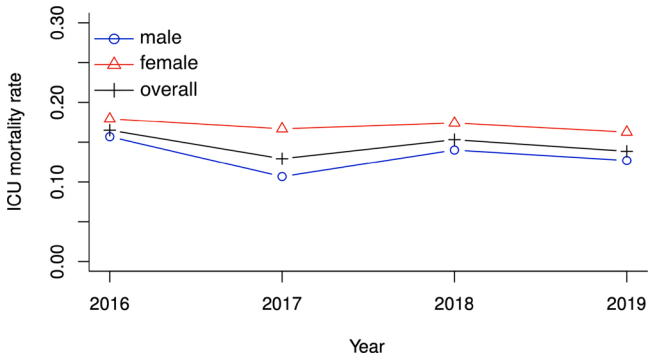

Fig. 2. Evolution of mortality rate at the ICU with year, disaggregated by sex, and for the overall population.

Appendix B. Tables for output *Result*

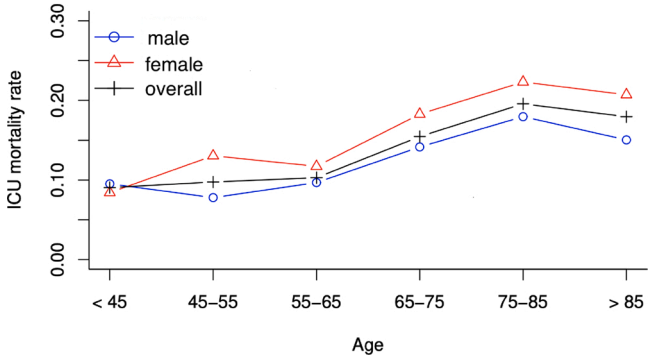

Fig. 3. Evolution of mortality rate at the ICU with age, disaggregated by sex, and for the overall population.

Appendix C. Tables for output *Destination*

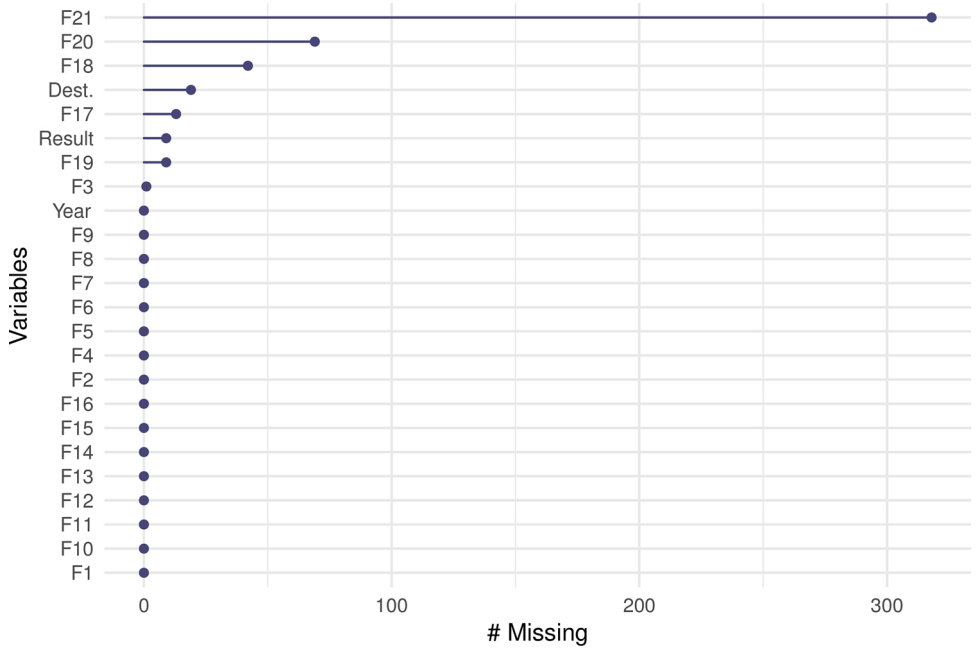

Fig. 4. Variables ordered by the number of missing values. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

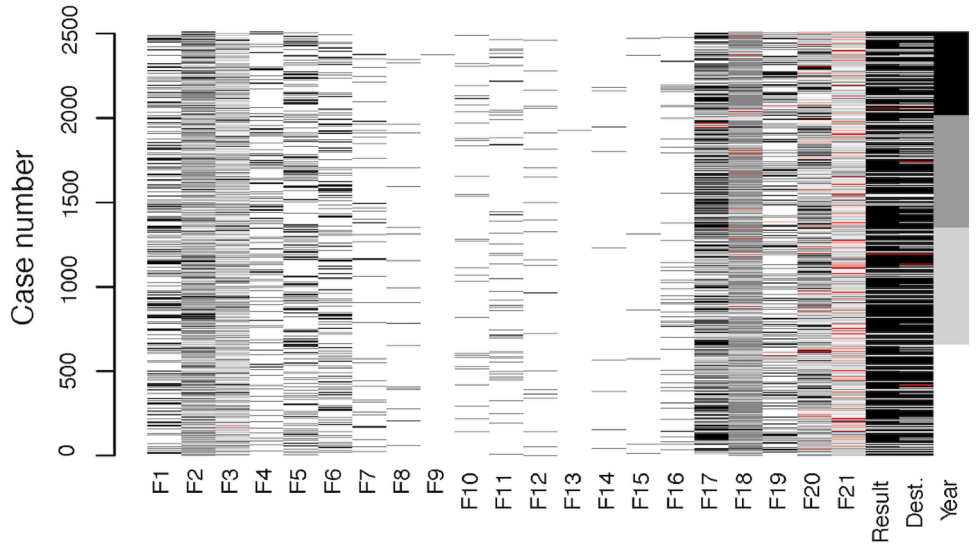

Fig. 5. Distribution of missing values (in red), where cases have been ordered by year. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

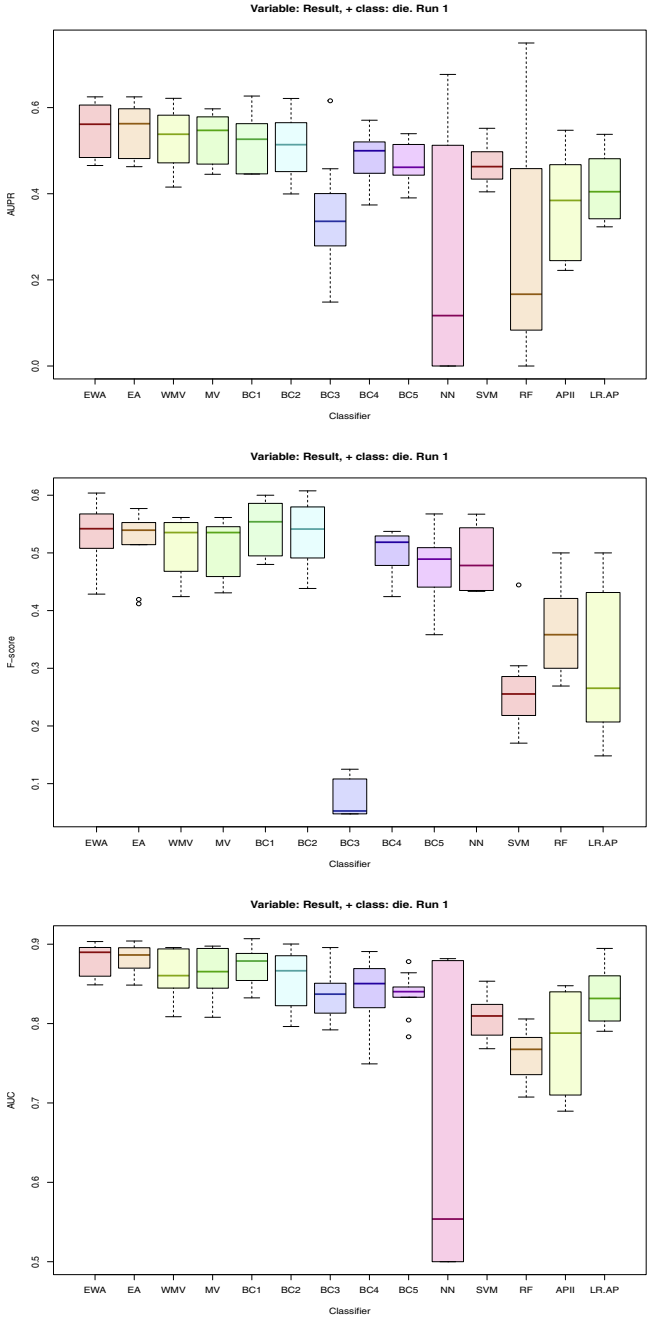

Fig. 6. Boxplots for output variable *Result* in the first run. AUPR, F-score and AUC. F-score cannot be computed for the standard approach based on APACHE II ("APII" as *x*-axis label), but it can for the locally recalibrated LR.APACHEII ("LR.AP" as *x*-axis label).

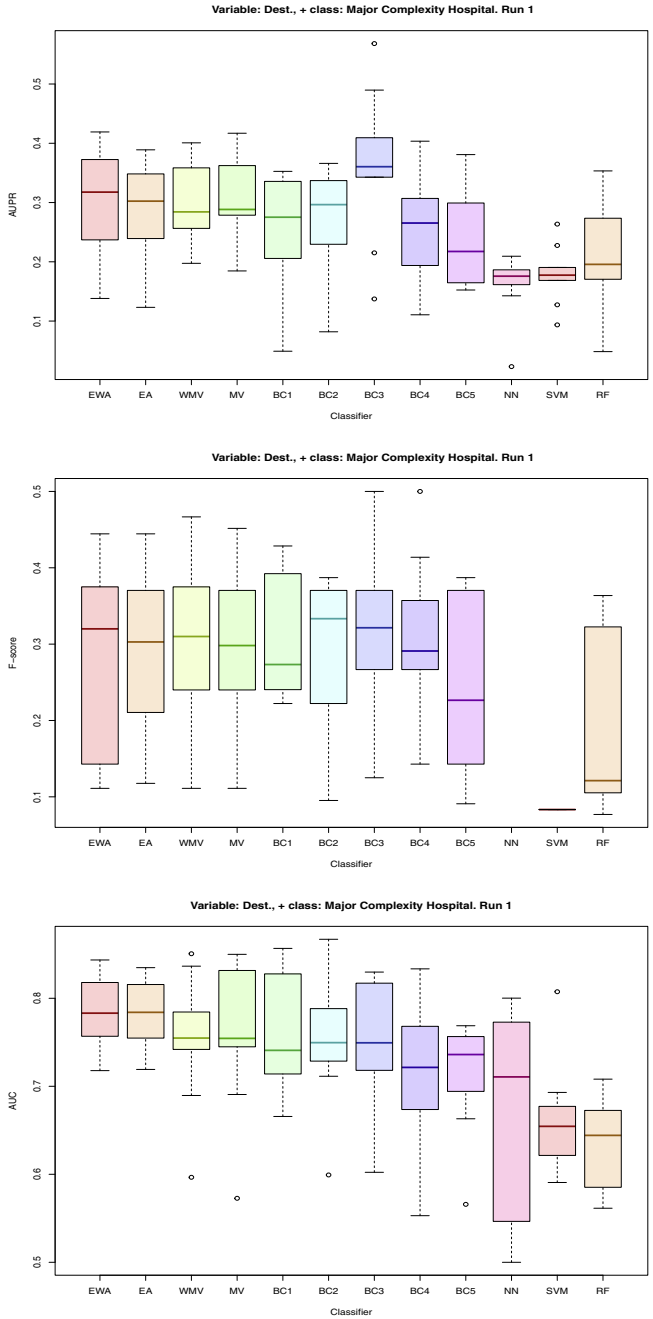

Fig. 7. Boxplots for output variable *Destination* in the first run, for AUPR, F-score and AUC.

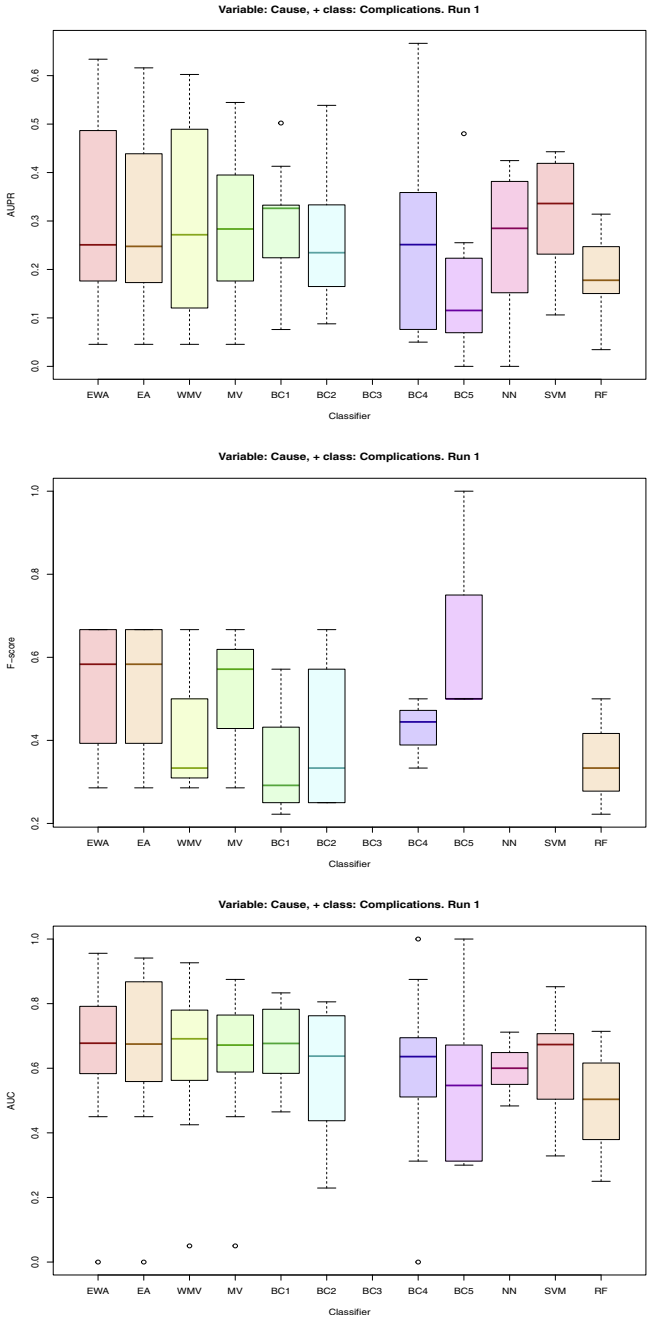

Fig. 8. Boxplots for output variable *Cause* in the first run, for AUPR, F-score and AUC.

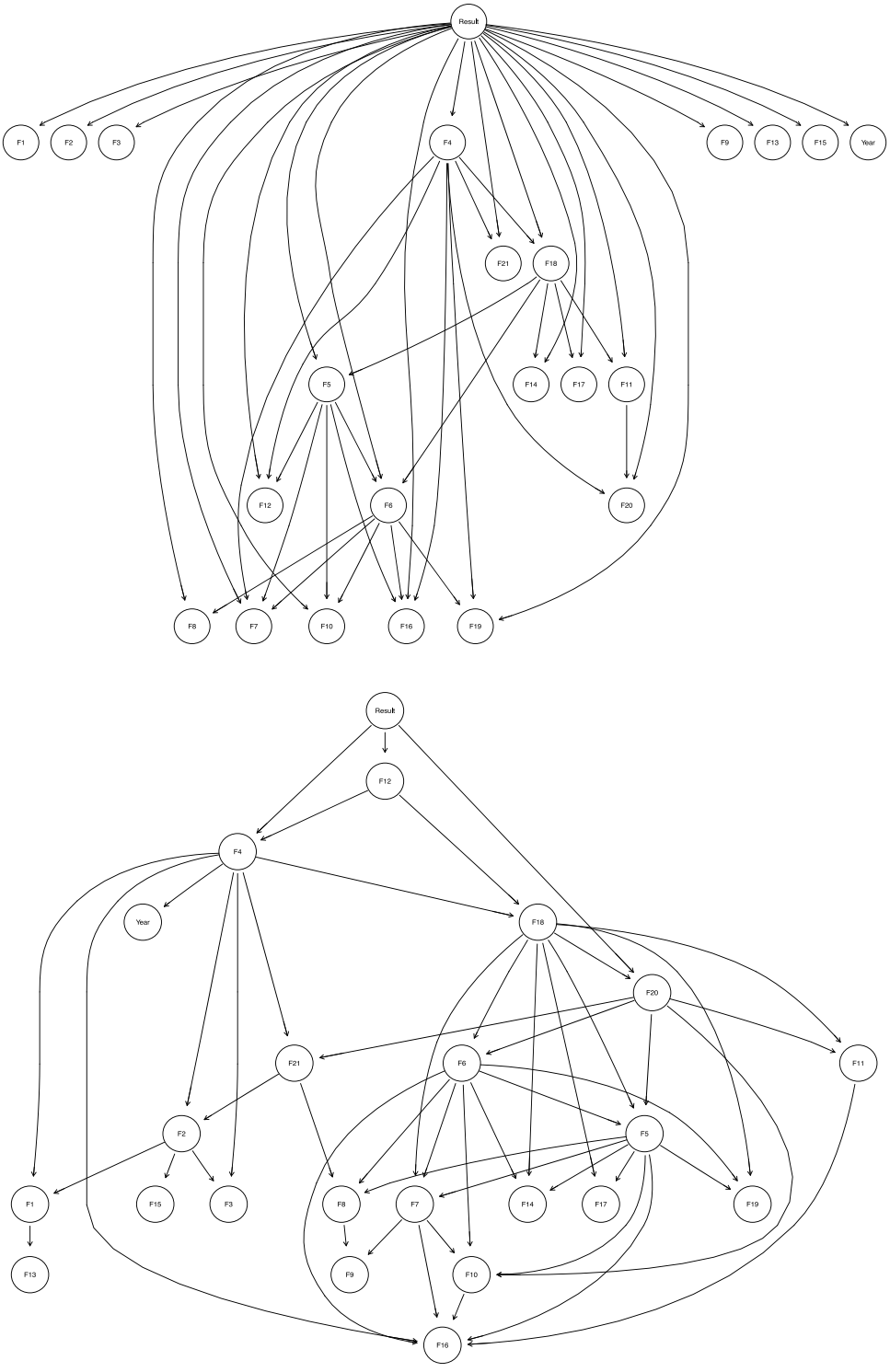

Fig. 9. DAGs for the base classifiers BC_2 (top) and BC_3 (bottom), learned from the whole database set.

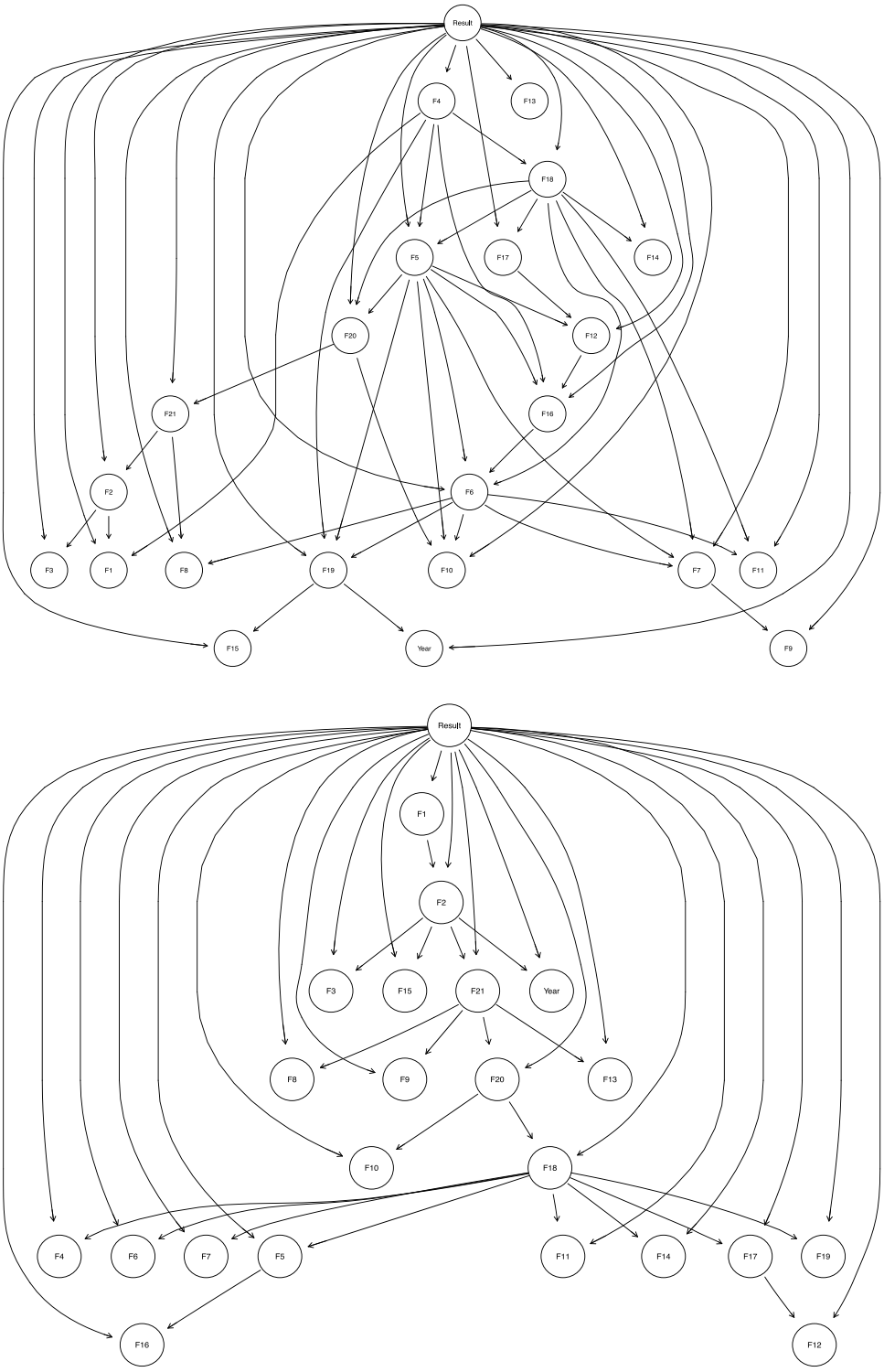

Fig. 10. DAGs for the base classifiers BC_4 (top) to BC_5 (bottom), learned from the whole database set.

Table 17
AUPR median for EWA/EA is significantly greater (p-value<0.1) than that of WMV and/or MV, for output variable *Result* and for each run? ("2" if it is greater for both, WMV and MV, "1" if it is greater for only one, "0" if it is not greater for either of them).

AUPR	Run																			
Result	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
EWA	0	0	1	2	2	0	2	0	2	1	1	2	2	0	1	1	1	1	2	2
EA	0	0	0	2	2	0	2	0	2	1	1	2	1	0	0	1	1	0	2	1

Table 18
Adjusted p-values for the comparisons between the four ensembles, corresponding to the statistical significances in Table 17 when we compare EWA/EA against WMV and MV. Also, in boldface, the adjusted p-values corresponding to the comparison between the two EWA and EA, that were not reported there.

AUPR (Result)	EWA	EA		WMV		MV	
			0.0059**		0.029*		0.0342*
			0.029*		0.018*		0.098·
		(5)	**0.0059****		0.024*		0.018*
			0.0342*		0.0117*		0.018*
			0.049*	**(8)**	0.039*		0.0098**
					0.029*		0.082·
					0.012*		0.012*
EWA>					0.055·	**(15)**	0.056·
							0.074·
							0.0059**
							0.029*
							0.093·
							0.059·
							0.012*
							0.018*
					0.049*		0.098·
					0.029*		0.027*
					0.056·		0.039*
				(8)	0.0645·		0.0146*
EA>					0.024*	**(10)**	0.093·
					0.049*		0.068·
					0.012*		0.012*
					0.018*		0.034*
							0.029*
							0.021*

Table 22
Adjusted p-values for the comparisons between the four ensembles, corresponding to the statistical significances in Table 21 when we compare EWA/EA against WMV and MV. Also, in boldface, the adjusted p-values corresponding to the comparison between the two EWA and EA, that were not reported there.

F-score (Result)	EWA	EA		WMV		MV	
			0.054·		0.027*		0.027*
		(2)	**0.0391***		0.027*		0.034*
				(5)	0.0977·		0.029*
					0.066·		0.0059**
EWA>					0.0391*	**(9)**	0.027*
							0.059·
							0.0059**
							0.082·
							0.018*
				(1)	0.0645·		0.027*
							0.049*
EA>						**(5)**	0.0146*
							0.032*
							0.0146*
WMV>						**(2)**	**0.0907**·
							0.090·

Table 19
Non-adjusted p-values corresponding to the comparisons between EWA and EA in Table 18, but only between them two (so the p-values are not adjusted). In boldface the 5 runs corresponding to the adjusted p-values that have been reported in Table 18.

AUPR (Result)	Run	**3**	**5**	**9**	14	**15**	18	**19**	20
EWA>EA	p-Value	0.00098***	0.014*	0.00098***	0.042*	0.0068**	0.08·	0.024*	0.042*

Table 20
(Non-adjusted) p-values corresponding to the comparison between LR.APACHEII and the standard based on the APACHE II score, for the alternative hypothesis that the former has greater AUPR median.

AUPR	Run	2	3	4	5	7	8	9	10	11
	p-Value	0.042*	0.0049**	0.014*	0.032*	0.0068**	0.0098**	0.0049**	0.0098**	0.032*
(Result)	Run	12	13	14	15	16	17	18	19	20
	p-Value	0.042*	0.065·	0.053·	0.0068**	0.08·	0.0049**	0.024*	0.065·	0.014*

Table 21
F-score median for EWA/EA is significantly greater (p-value<0.1) than that of WMV and/or MV, for output variable *Result* and for each run? ("2" if it is greater for both, WMV and MV, "1" if it is greater for only one, "0" if it is not greater for either of them).

F-score	Run																			
Result	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
EWA	0	0	0	0	2	0	2	0	0	1	2	2	0	1	0	0	2	1	0	1
EA	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0	2	0	0	0

Table 23
Non-adjusted *p*-values corresponding to the comparisons between EWA and EA in Table 22, but only between them two (so the *p*-values are not adjusted). In boldface the 2 runs corresponding to the adjusted *p*-values that have been reported in Table 22.

F-score (Result)	Run	1	2	3	**5**	8	9	10
	p-Value	0.062·	0.078·	0.026*	0.018*	0.029*	0.071·	0.038*
EWA>EA	Run	11	13	14	**17**	18	19	20
	p-Value	0.062·	0.054·	0.022*	0.012*	0.029*	0.030*	0.029*

Table 24
AUC median for EWA/EA is significantly greater (*p*-value<0.1) than that of WMV and/or MV, for output variable *Result* and for each run? ("2" if it is greater for both, WMV and MV, "1" if it is greater for only one, "0" if it is not greater for either of them).

AUC	Run																			
Result	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
EWA	2	0	2	2	2	0	2	2	2	2	0	2	2	0	2	2	0	0	2	2
EA	2	0	2	2	2	0	2	2	2	2	0	2	2	0	2	2	0	0	2	2

Table 25
Non-adjusted *p*-values corresponding to the comparisons between EWA and EA when comparing them alone (non-adjusted *p*-values), for the AUC metric and the output variable *Result*.

AUC (Result)	Run	1	2	3	4	5	9	10	14	15	18	19	20
EWA>EA	*p*-Value	0.08·	0.024*	0.0029**	0.096·	0.002**	0.0068**	0.0098**	0.08·	0.012*	0.065·	0.02*	0.065·

Table 26
Adjusted *p*-values for the comparisons between the four ensembles, corresponding to the statistical significances in Table 24 when we compare EWA/EA against WMV and MV. Also, in boldface, the adjusted *p*-values corresponding to the comparison between the two EWA and EA, that were not reported there, indicating to which run they correspond.

AUC (Result)	EWA	EA		WMV		MV	
EWA>		**(6)**	**0.018** * (run 3)	**(14)**	0.029*	**(14)**	0.029*
			0.012 * (run 5)		0.041*		0.023*
			0.034 * (run 9)		0.018*		0.018*
			0.059 · (run 10)		0.049*		0.049*
			0.039 * (run 15)		0.012*		0.020*
			0.012 * (run 19)		0.074·		0.034*
					0.029*		0.034*
					0.074·		0.059·
					0.029*		0.029*
					0.012*		0.012*
					0.039*		0.029*
					0.018*		0.012*
					0.012*		0.015*
					0.082·		0.082·
EA>				**(14)**	0.039*	**(14)**	0.039*
					0.049*		0.023*
					0.020*		0.020*
					0.049*		0.049*
					0.012*		0.029*
					0.097·		0.029*
					0.034*		0.034*
					0.084·		0.074·
					0.029*		0.029*
					0.012*		0.012*
					0.039*		0.029*
					0.012*		0.041*
					0.012*		0.015*
					0.082·		0.082·

Table 27
(Non-adjusted) *p*-values corresponding to the comparison between LR.APACHEII and APACHEII, for the alternative hypothesis that the former has greater AUPR median.

AUPR	Run	2	3	4	5	7	8	9	10	11
	p-Value	0.042*	0.0049**	0.014*	0.032*	0.0068**	0.0098**	0.0049**	0.0098**	0.032*
(Result)	Run	12	13	14	15	16	17	18	19	20
	p-Value	0.042*	0.065·	0.053·	0.0068**	0.08·	0.0049**	0.024*	0.065·	0.014*

Table 28
Adjusted *p*-values for the comparisons between the four ensembles. Only significant *p*-values (<0.1) have been recorded.

AUPR (Destination)	EA	
EWA>	**(10)**	0.041* 0.0059** 0.018* 0.012* 0.012* 0.018* 0.012* 0.082· 0.082· 0.082·
WMV>	**(2)**	0.082· 0.059·

Table 29
Non-adjusted *p*-values corresponding to the comparisons between EWA and EA in Table 28, for output variable *Destination* and AUPR. In boldface the 10 runs for which there is statistical significance when considering adjusted *p*-values for comparison of the four ensembles in Table 28, for comparison between EWA and EA.

AUPR (Destination)	Run	1	**2**	3	**4**	**5**	**6**	**7**	**8**	9	10
	p-Value	0.019*	0.0068**	0.042*	0.00098***	0.0029**	0.002**	0.002**	0.0029**	0.053·	0.024*
EWA>EA	Run	**11**	**12**	13	14	**15**	16	17	18	**20**	
	p-Value	0.002*	0.014*	0.042*	0.019*	0.014*	0.032*	0.019*	0.042*	0.014*	

Table 30
Number of runs, say n, for which AUPR of BC_3 model (for the output variable *Destination*) is statistically greater than that of the other classifiers, when considering adjusted *p*-values for comparison of five classifiers at the same time (BC_1 to BC_5 first, and secondly BC_3 and the ensembles). Below appear the corresponding *p*-values for the exact Binomial test in favour of BC_3, which are 0.5^n.

AUPR (Destination)	BC_1	BC_2	BC_4	BC_5	EWA	EA	WMV	MV
	9	10	16	20	6	10	8	7
BC_3>	0.002**	0.0009***	1.53×10^{-5}***	9.54×10^{-7}***	0.016*	0.0009***	0.004**	0.008**

Table 31
(Non-adjusted) *p*-values corresponding to the comparison between EWA and EA for output variable *Destination* and F-score.

F-score (Destination)	Run	3	8	10	11	13	14	15
EWA>EA	*p*-Value	0.03*	0.05·	0.09·	0.05·	0.05·	0.02*	0.08·

Table 32
Number of runs for which there are statistically significant differences between EWA and each of the classifiers EA, BC_1, BC_2 and BC_3, for the output variable *Destination* and the metric AUC. These differences are always in the sense that the median of EWA is greater than that of the others. The one-sided *p*-values for the exact Binomial test for the statistical significance of the number of runs is also given in the second column.

AUC (Destination)	Number of runs	*p*-Value
EWA>EA	15	$0.5^{15} = 3.05 \times 10^{-5}$***
EWA>BC_1	7	$0.5^{7} = 0.0078$**
EWA>BC_2	9	$0.5^{9} = 0.0020$**
EWA>BC_3	4	$0.5^{4} = 0.0625$·

Appendix D

D.1 Coefficients (weights) β for Eq. (2)

Table 33
Sepsis means $F_{19} = 1$. Non surgical category means F_{18} = Coronary, Medical or Trauma, while surgical category means F_{18} = Elective or Urgent Surgical. Blanc spaces mean excluding category. This table has been adapted to our setting from [9].

Features	Sepsis or non surgical category	No sepsis and surgical category
F_4	−0.191	−0.797
F_5	−0.890	−0.610
F_6	0.493	−0.797
F_7	−0.759	−1.150
F_8	−0.885	−0.196
F_9	0.501	−0.613
F_{10}	0.393	0.393
F_{11}		−0.248
F_{12}	−1.368	−0.797
F_{13}	−0.517	−0.955
F_{14}	−1.228	−1.684
F_{15}	−0.142	−0.196
F_{19}	0.113	

D.2 Coefficients (weights) α for Eq. (3)

Table 34
Coefficients α for Eq. (3) (only those with significant p-values, that is >0.10, have been recorded). *(a)*: the odds in favour of "die" where the regressors are at their reference value (all equal to "0", including APACHE II). *(b)*: increase in odds in favour of "die" for a one-unit increase in APACHE II score, holding the other regressors at a fixed value. *(c)*: increase in odds in favour of "die" for the regressor taken the value "1", with respect to value "0", holding the other regressors at a fixed value. *(d)*: decrease in odds in favour of "die" for F_4 taken the value "1", with respect to value "0", holding the other regressors at a fixed value.

Features	α estimated	p-Value	interpretation
Intercept	$\alpha_0 = -4.85711$	5.72×10^{-14}***	$0.00777^{(a)}$
APACHE II	$\alpha_1 = 0.11544$	$<2 \times 10^{-16}$***	12.2 % (b)
F_{19}	$\alpha_3 = 0.48604$	0.00212**	62.6 % (c)
F_4	$\alpha_4 = -1.83024$	0.03768*	**84.1 % (d)**
F_5	$\alpha_5 = 0.62866$	0.00126**	87.5 % (c)
F_6	$\alpha_6 = 0.52522$	0.00896**	69.1 % (c)
F_7	$\alpha_7 = 0.49130$	0.05651·	63.4 % (c)
F_{10}	$\alpha_{10} = 1.82376$	4.38×10^{-9}***	519.5 % (c)

Appendix E. Centrality and betweenness measures

Table 35
(Normalized to sum up 100) Freeman's degree of centrality and Basic standard betweenness measure of the features. In boldface the highest 5 values of each column.

Feature	Freeman's centrality (%)				Basic standard betw. (%)			
	BC_2	BC_3	BC_4	BC_5	BC_2	BC_3	BC_4	BC_5
F_1	1.5	2.0	2.0	1.5	0.0	6.1	0.0	0.0
F_2	1.5	3.0	4.5	**10.0**	0.0	**13.0**	7.0	0.0
F_3	1.5	2.0	2.0	3.5	0.0	0.0	0.0	**14.5**
F_4	**12.0**	**9.5**	**6.5**	3.5	0.0	**15.6**	0.0	0.0
F_5	**10.5**	**10.5**	**11.0**	**5.0**	**16.7**	6.9	7.0	0.0
F_6	**12.0**	**9.5**	**10.0**	3.5	**38.9**	4.8	**23.3**	**3.6**
F_7	**6.0**	6.5	5.5	3.5	0.0	4.7	5.5	0.0
F_8	3.0	4.0	3.5	3.5	0.0	3.1	0.0	0.0
F_{10}	4.5	5.5	4.5	3.5	0.0	0.2	0.0	0.0
F_{11}	4.5	3.0	3.5	3.5	**5.6**	0.4	0.0	0.0
F_{12}	4.5	3.0	4.5	3.5	0.0	1.5	7.8	0.0
F_{16}	**6.0**	6.5	5.5	3.5	0.0	0.0	**14.6**	0.0
F_{17}	3.0	2.0	3.5	**5.0**	0.0	0.0	0.4	0.0
F_{18}	**10.5**	**10.5**	**10.0**	**16.5**	**38.9**	**19.6**	5.2	**3.6**
F_{19}	4.5	3.0	**6.5**	3.5	0.0	0.0	**10.9**	**29.0**
F_{20}	4.5	**7.5**	5.5	**5.0**	0.0	**12.6**	**8.9**	0.0
F_{21}	3.0	4.0	4.5	**10.0**	0.0	**11.4**	**9.4**	**26.1**

Table 36
(Normalized to sum up 100) Borgatti's proximal source and proximal target betweenness measures of the features. In boldface the highest 5 values of each column.

Feature	Borgatti's proximal source (%)				Borgatti's proximal target (%)			
	BC_2	BC_3	BC_4	BC_5	BC_2	BC_3	BC_4	BC_5
F_1	0.0	**8.5**	0.0	0.0	0.0	2.4	0.0	0.0
F_2	0.0	**14.6**	**12.5**	0.0	0.0	**6.1**	2.8	0.0
F_3	0.0	0.0	0.0	**6.7**	0.0	0.0	0.0	**33.3**
F_4	0.0	**14.6**	0.0	0.0	0.0	**21.8**	0.0	0.0
F_5	**18.75**	**8.8**	**7.9**	0.0	**15.6**	6.7	**12.5**	0.0
F_6	**43.75**	6.0	**27.5**	**8.3**	**34.4**	4.2	**17.8**	**1.7**
F_7	0.0	6.6	**9.7**	0.0	0.0	3.0	4.2	0.0
F_8	0.0	4.4	0.0	0.0	0.0	2.4	0.0	0.0
F_{10}	0.0	0.3	0.0	0.0	0.0	0.3	0.0	0.0
F_{11}	**6.25**	0.6	0.0	0.0	**6.25**	0.6	0.0	0.0
F_{12}	0.0	0.6	1.4	0.0	0.0	2.1	**13.9**	0.0
F_{16}	0.0	0.0	3.2	0.0	0.0	0.0	**13.4**	0.0
F_{17}	0.0	0.0	0.7	0.0	0.0	0.0	0.7	0.0
F_{18}	**31.25**	**22.2**	6.0	**8.3**	**43.75**	**24.6**	9.3	**1.7**
F_{19}	0.0	0.0	**19.4**	**53.3**	0.0	0.0	8.3	**16.7**
F_{20}	0.0	7.6	4.6	0.0	0.0	**17.0**	**11.6**	0.0
F_{21}	0.0	5.1	6.9	**10.0**	0.0	**8.6**	5.6	**20.0**

Appendix F. CPT tables for the output *Result*

Each table is obtained computing the *a posteriori* conditional probability of the output *Result* to the feature, assuming the rest of features are not observed, except in the case of the "Main cause of admission" category, for which when one is present, the others are necessarily absent, since thery are mutually exclusive[4] ; instead, if it is absent, we assume the others are not observed.

Table 37
CPT of variable *Result* conditioned to F_1 and to F_2.

	F_1: Sex		F_2: Age					
	Male	Female	<45	45–54	55–64	65–74	75–84	>84
live	**0.86575**	**0.83143**	**0.90517**	**0.89967**	**0.89313**	**0.84506**	**0.80894**	**0.82340**
die	0.13425	0.16857	0.09483	0.10033	0.10687	0.15494	0.19106	0.17660

Table 38
CPT of variable *Result* conditioned to F_3 and to F_4.

	F_3: Charlson comorbidity index					F_4: ACS	
	0	1	2	3	>3	0	1
live	**0.90627**	**0.86975**	**0.84657**	**0.81614**	**0.76523**	**0.82243**	**0.99610**
die	0.09373	0.13025	0.15343	0.18386	0.23477	0.17757	0.00390

Table 39
CPT of variable *Result* conditioned to F_5, to F_6, to F_7 and to F_8.

	F_5: RF		F_6: Shock		F_7: Coma		F_8: Renal F	
	0	1	0	1	0	1	0	1
live	**0.87220**	**0.83919**	**0.88342**	**0.79644**	**0.86008**	**0.83497**	**0.85822**	**0.84957**
die	0.12780	0.16081	0.11658	0.20356	0.13992	0.16503	0.14178	0.15043

Table 40
CPT of variable *Result* conditioned to F_9, to F_{10}, to F_{11} and to F_{12}.

	F_9: Hepatic F		F_{10}: CRA		F_{11}: ES		F_{12}: Arrhythmia	
	0	1	0	1	0	1	0	1
live	**0.85302**	**0.98666**	**0.87396**	0.37702	**0.84377**	**0.99580**	**0.85084**	**0.94933**
die	0.14698	0.01334	0.12604	**0.62298**	0.15623	0.00420	0.14916	0.05067

Table 41
CPT of variable *Result* conditioned to F_{13}, to F_{14}, to F_{15} and to F_{16}.

	F_{13}: CT		F_{14}: OT		F_{15}: Intoxication		F_{16}: Other syndromes	
	0	1	0	1	0	1	0	1
live	**0.85369**	**0.69640**	**0.85177**	**0.97827**	**0.85271**	**0.92872**	**0.84672**	**0.96777**
die	0.14631	0.30360	0.14823	0.02173	0.14729	0.07128	0.15328	0.03223

Table 42
CPT of variable *Result* conditioned to F_{17}.

	F_{17}: Origin					
	Ward	Operation Room	Extra Hospital Emergency	Other Hospital	Emergency Room	unknown
live	**0.77059**	**0.89578**	**0.53767**	**0.87632**	**0.88156**	**0.71747**
die	0.22941	0.10422	0.46233	0.12368	0.11844	0.28253

[4] Although it is possible for a patient to present more than one of the features of the "Main cause of admission" category, in practice, there were cases in which several were recorded, in principle only the most significant had to be reported and, in fact, this is so for almost 89% of patients.

Table 43
CPT of variable *Result* conditioned to F_{18}.

	F_{18}: Generic Syndrome					
	Elective surgical	Urgent surgical	Coronary	Medical	Trauma	unknown
live	**0.94701**	**0.85312**	**0.97963**	**0.80451**	**0.93671**	**0.93971**
die	0.05299	0.14688	0.02037	0.19549	0.06329	0.06029

Table 44
CPT of variable *Result* conditioned to F_{19}.

	F_{19}: Sepsis	
	No	Yes
live	**0.89398**	**0.78065**
die	0.10602	0.21935

Table 45
CPT of variable *Result* conditioned to F_{20}.

	F_{20}: ICU Workload					
	M. monitoring	M. unstable without coma/shock	M. unstable coma/shock	Post-surg. monitor.	Post-surg. unstable	Unknown
live	**0.95564**	**0.86603**	**0.55982**	**0.99195**	**0.98728**	**0.85176**
die	0.04436	0.13397	0.44018	0.00805	0.01272	0.14824

Table 46
CPT of variable *Result* conditioned to F_{21}.

	F_{21}: APACHE II (discretized)								
	<5	5–9	10–14	15–19	20–24	25–29	30–34	>34	Unknown
live	**0.99579**	**0.96505**	**0.92690**	**0.86155**	**0.71854**	**0.59862**	**0.58245**	0.36033	**0.76492**
die	0.00421	0.03495	0.07310	0.13845	0.28146	0.40138	0.41755	**0.63967**	0.23508

References

[1] Kerlin MP, Cooke CR. Understanding costs when seeking value in critical care. Ann Am Thorac Soc 2015;12(12):1743–4.
[2] Lone NI, Gillies MA, Haddow C, Dobbie R, Rowan KM, Wild SH, et al. Five-year mortality and hospital costs associated with surviving intensive care. Am J Respir Crit Care Med 2016;194(2):198–208.
[3] Detsky ME, Harhay MO, Bayard DF, Delman AM, Buehler AE, Kent SA, et al. Six-month morbidity and mortality among intensive care unit patients receiving life-sustaining therapy. A prospective cohort study. Ann Am Thorac Soc 2017;14(10):1562–70.
[4] Granholm A, Miller MH, Krag M, Perner A, Hjortrup PB. Predictive performance of the Simplified Acute Physiology Score (SAPS) II and the initial Sequential Organ Failure Assessment (SOFA) score in acutely ill intensive care patients: post-hoc analyses of the SUP-ICU inception cohort study. PLOS ONE 2016;11(12):e0168948. 10.1371/journal.pone.0168948.
[5] Li Z, Cheng B, Wang J, Xie G, Yu X, Huang M, et al. A multifactor model for predicting mortality in critically ill patients: a multicenter prospective cohort study. J Crit Care 2017;42:18–24.
[6] McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, et al. The quality of health care delivered to adults in the United States. N Engl J Med 2003;348(26):2635–45.
[7] Steinberg EP. Improving the quality of care. Can we practice what we preach? N Engl J Med 2003;348(26):2681–3.
[8] Niewiński G, Starczewska M, Kański A. Prognostic scoring systems for mortality in intensive care units. The APACHE model. Anaesthesiol Intensive Ther 2014;46(1):46–9.
[9] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985;13(10):818–29.
[10] Theresa SJ, Latheef F. Evaluation of acute physiology and chronic health evaluation (APACHE) II in predicting ICU mortality among critically ill. Int J Adv Med 2017;4(6):1566–72.
[11] Sekuli AD, Trpkovic SV, Pavlovic AP, Marinkovic OM, Llic AN. Scoring systems in assessing survival of critically Ill ICU patients. Med Sci Monit 2015;21:2621–9.
[12] Godinjal A, Iglica A, Rama A, Tancica I, Jusufovic S, Ajanovic A, et al. Predictive value of SAPS II and APACHE II scoring systems for patient outcome in a medical intensive care unit. Acta Med Acad 2016;45(2):97–103.
[13] Barado Barado J, Guergué JM, Esparza L, Azcárate C, Mallor F, Ochoa S. A mathematical model for simulating daily bed occupancy in an intensive care unit. Crit Care Med 2012;40(4):1098–104.
[14] Garg AX. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293:1223–38.
[15] Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006;144:742–52.
[16] Tu JV, Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Comp Biomed Res 1993;26:220–9.
[17] Doig G, Inman K, Sibbald W, Martin C, Robertson J. Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. In: Proceedings of the annual symposium on computer application in medical care, 1993; 1993. p. 361–5.
[18] Buchman TG, Kubos KL, Seidler AJ, et al. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care uni. Crit Care Med 1994;22:750–62.
[19] Dybowski R, Gant V, Weller P, Chang R. Prediction of outcome in critically ill patients using artificial neural networks synthesised by genetic algorithm. Lancet 1996;347:1146–50.
[20] Hanson III CW, Marshall BE. Artificial intelligence applications in the intensive care unit. Crit Care Med 2001;29:427–35.
[21] Luaces O, Taboada F, Albaiceta GM, Domínguez LA, Enríquez P, Behamonde A. Predicting the probability of survival in intensive care unit patients from a small number of variables and training examples. Artif Intell Med 2009;45:63–76.
[22] Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 2018;83:112–34.
[23] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.
[24] Sadeghi R, Banerjee T, Romine W. Early hospital mortality prediction using vital signals. Submitted to smart health. 2019 [cs.LG] 9 Feb 2019, arXiv:1803.06589v2.
[25] Caicedo-Torres W, Gutiérrez J. ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU. 2019 [cs.LG] 24 Jan 2019, arXiv: 1901.08201v1.

[26] Overweg H, Popkes AL, Ecole A, Li Y, Hernández-Lobato JM, Zaykov Y, et al. Interpretable outcome prediction with sparse bayesian neural networks in intensive care. 2019 [cs.LG] 9 Sep 2019, arXiv:1905.02599v2.
[27] Maas AIR, Menon DK, Steyerberg EW, Citerio G, Lecky F, Manley GT, et al. Collaborative european neurotrauma effectiveness research in traumatic brain injury (center-tbi): a prospective longitudinal observational study. Neurosurgery 2014;76(1):67–80.
[28] Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med 2015;3:42–52.
[29] Aczon M, Ledbetter D, Ho L, Gunny A, Flynn A, Williams J, et al. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. 2017. Cs Math Q-Bio Stat 2017 Jan 23, arXiv:170106675.
[30] Spiegelhalter DJ. Incorporating Bayesian ideas into healthcare evaluation. Stat Sci 2004;19:156–74.
[31] Walshe T, Burgman M. A framework for assessing and managing risks posed by emerging diseases. Risk Anal 2010;30(2):236–49.
[32] Cruz-Ramírez N, Acosta-Mesa HG, Carrillo-Calvet H, Alonso Nava-Fernández L, Barrientos-Martínez RE. Diagnosis of breast cancer using BN: a case study. Comput Biol Med 2007;37:1553–64.
[33] Gade J, Rosenfalck A, van Gils M, et al. Modelling techniques and their application for monitoring in high dependency environments-learning models. Comput Methods Programs Biomed 1996;51:75–84.
[34] Nikiforidis GC, Sakellaropoulos GC. Expert system support using Bayesian belief networks in the prognosis of head-injured patients of the ICU. Med Inf 1998;23:1–18.
[35] Sandri M, Berchialla P, Baldi I, Gregori D, De Blasi RA. Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU. J Biomed Inform 2014;48:106–13.
[36] Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinial outomes for patients with sepsis: a machine learning basis for a clinical decision support system. J Am Med Inform Assoc 2014;21:315–25.
[37] Delgado R, Núñez-González JD, Yébenes JC, Lavado A. Vital prognosis of patients in intensive care units using an Ensemble of Bayesian Classifiers. In: Proceedings of the LOD 2019; 2019 [to appear in].
[39] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29(2–3):131–63.
[40] Davis J, Goadrich M. The relationship between precision-recall and roc curves.. In: Proceedings of the 23rd international conference on machine learning; 2006. p. 233–40.
[41] He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng 2009;21(9):1263–84.
[42] Khadanga S, Aggarwal K, Joty S, Srivastava J. Using clinical notes with time series data for ICU management. 2019 [cs.CL] 12 Sep 2019, arXiv:1909.09702v1.
[43] Scutari M. Learning Bayesian Networks with the bnlearn R Package. J Stat Softw 2010;35(3):1–22.
[44] Hojsgaard S. Graphical independence networks with the gRain package for R. J Stat Softw 2012;46(10):1–26.
[45] Wilcoxon F. Individual comparisons by ranking methods. Biometr Bull 1945;1(6):80–3.
[46] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52(3–4):591–611.
[47] Delgado R, Gonzalez JL, Sotoca A, Tibau XA. Archetypes of wildfire arsonists: an approach by using bayesian networks. Forest Fire Cap 2018;2:25–50. Janusz Szmyt, IntechOpen.
[48] Freeman LC. A set of measures of centrality based upon betweenness. Sociometry 1977;40:35–41.
[49] Delgado R, Tibau XA, et al. Measuring features strength in probabilistic classification. In: Medina J, editor. Information processing and management of uncertainty in knowledge-based systems. Theory and foundations. IPMU 2018. Communications in computer and information science, vol. 853. Cham: Springer; 2018.