Feature selection for the accurate prediction of septic and cardiogenic shock ICU mortality in the acute phase

Aushev, Alexander; Ripoll, Vicent Ribas; Vellido, Alfredo; Aletti, Federico; Bollen Pinto, Bernardo; Herpain, Antoine; Post, Emiel Hendrik; Medina, Eduardo Romay; Ferrer, Ricard; Baselli, Giuseppe; Bendjelid, Karim

How to cite

AUSHEV, Alexander et al. Feature selection for the accurate prediction of septic and cardiogenic shock ICU mortality in the acute phase. In: PLOS ONE, 2018, vol. 13, n° 11, p. e0199089. doi: 10.1371/journal.pone.0199089

This publication URL: https://archive-ouverte.unige.ch/unige:115362

Publication DOI: 10.1371/journal.pone.0199089

RESEARCH ARTICLE

Feature selection for the accurate prediction of septic and cardiogenic shock ICU mortality in the acute phase

Alexander Aushev[1☯], Vicent Ribas Ripoll[2☯]*, Alfredo Vellido[1☯], Federico Aletti[3], Bernardo Bollen Pinto[4,5], Antoine Herpain[6], Emiel Hendrik Post[6], Eduardo Romay Medina[7], Ricard Ferrer[8,9], Giuseppe Baselli[10], Karim Bendjelid[4,5]

1 Intelligent Data Engineering and Artificial Intelligence (IDEAI) Research Center, Computer Science, Universitat Politècnica de Catalunya, Barcelona, Spain, **2** Eurecat, Centre Tecnològic de Catalunya, eHealth, Data Analytics in Omics, Barcelona, Spain, **3** Department of Bioengineering, University of California San Diego, La Jolla, CA, United States of America, **4** Hemodynamic Research Group, Université de Genève, Geneva, Switzerland, **5** Anesthesiology Department, Hôpitaux Universitaires de Genève, Geneva, Switzerland, **6** Department of Intensive Care, Hôpital Erasme - Université Libre de Bruxelles, Brussels, Belgium, **7** Hospital Universitari Mutua de Terrassa, Terrassa, Spain, **8** Critical Care Department, Vall d'Hebron University Hospital, Barcelona, Spain, **9** Shock, Organ Dysfunction and Resuscitation (SODIR) Research Group, Vall d' Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain, **10** Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano - Piazza Leonardo da Vinci, Milan, Italy

☯ These authors contributed equally to this work.
* vicent.ribas@eurecat.org

Abstract

Circulatory shock is a life-threatening disease that accounts for around one-third of all admissions to intensive care units (ICU). It requires immediate treatment, which is why the development of tools for planning therapeutic interventions is required to deal with shock in the critical care environment. In this study, the ShockOmics European project original database is used to extract attributes capable of predicting mortality due to shock in the ICU. Missing data imputation techniques and machine learning models were used, followed by feature selection from different data subsets. Selected features were later used to build Bayesian Networks, revealing causal relationships between features and ICU outcome. The main result is a subset of predictive features that includes well-known indicators such as the SOFA and APACHE II scores, but also less commonly considered ones related to cardiovascular function assessed through echocardiograpy or shock treatment with pressors. Importantly, certain selected features are shown to be most predictive at certain time-steps. This means that, as shock progresses, different attributes could be prioritized. Clinical traits obtained at 24h. from ICU admission are shown to accurately predict cardiogenic and septic shock mortality, suggesting that relevant life-saving decisions could be made shortly after ICU admission.

OPEN ACCESS

Citation: Aushev A, Ripoll VR, Vellido A, Aletti F, Pinto BB, Herpain A, et al. (2018) Feature selection for the accurate prediction of septic and cardiogenic shock ICU mortality in the acute phase. PLoS ONE 13(11): e0199089. https://doi.org/10.1371/journal.pone.0199089

Editor: Bobak Mortazavi, Texas A&M University, UNITED STATES

Received: May 25, 2018

Accepted: November 5, 2018

Published: November 20, 2018

Data Availability Statement: The code for this study can be found via the following GitHub URL: https://github.com/alexaushev/feature-selection-for-shock. Data are available from the ShockOmics Institutional Data Access for researchers who meet the criteria for access to confidential data. Our dataset contains sensitive information because patients may be potentially identified. The policies for data custody and anonymization were imposed and approved by all ethical committees in the three ICUs involved in our study: the Geneva Regional Research Ethics Committee (study number 14-

041), the Ethical Committee of Hopital Erasme-Universite Libre de Bruxelles (study number P2014/171), and the Mutua de Terrassa Hospital Institutional Review Board (study number EO/1407). For this reason, the Shockomics project has appointed a data access committee for data sharing. In compliance with PLoS One policies for data sharing, all requests should be addressed to: Prof. Dr. Roberta Pastorelli; Head of Protein and Metabolite Biomarkers Unit, Department of Environmental Health Sciences, IRCCS-Istituto di Ricerche Farmacologiche "Mario Negri", Via La Masa 19, 20156 Milano, Italy; tel.: +39 0239014397; fax: +39 0239014735; e-mail: roberta.pastorelli@marionegri.it.

Funding: The following authors were funded by the Shockomics grant (number: 602706, website: http://www.shockomics.org/, https://cordis.europa.eu/project/rcn/109573_en.html): VR Ripoll, F Aletti, BB Pinto, K Bendjelid, A Herpain, EH Post, ER Medina, R Ferrer, G Baseli. The following consortium members were involved in the study design and data collection: C Conti, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano - Piazza Leonardo da Vinci, Milan, Italy; M Ferrario, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano - Piazza Leonardo da Vinci, Milan, Italy; C Barlassina, Dipartimento di Scienze della Salute, Università degli Studi di Milano, Milan, Italy; E De Oliveira, Parc Científic de Barcelona, Barcelona, Spain; R Pastorelli, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy; G Tedeschi, Fondazione Filarete, Milan, Italy; G Ristagno, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy; FS Taccone, Department of Intensive Care, Hôpital Erasme - Université Libre de Bruxelles, Brussels, Belgium; GW Schmid-Shönbein, Department of Bioengineering, University of California San Diego, La Jolla, USA; D De Backer, Department of Intensive Care, Hôpital Erasme - Université Libre de Bruxelles, Brussels, Belgium; G Baselli, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano - Piazza Leonardo da Vinci, Milan, Italy. A Vellido was partially funded by TIN2016-79576-R research project of the Spanish Ministry of Science, Innovation and Universities "Programa Estatal de I+D+i Orientada a los Retos de la Sociedad." The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Shock (or circulatory shock) is a life-threatening medical condition that requires immediate treatment. It is prevalent in the Intensive Care Unit (ICU) and a major concern in Critical Care in general. This condition occurs when the organs and tissues of the body do not receive enough blood and, as a result, cells see their oxygen and nutrients supply restricted so that organs become damaged. Hypotension, tissue hypoperfusion and hyperlactatemia are amongst the most common symptoms [1]. The outcome of an individual case depends on the stage of shock, the underlying condition and the general medical state of the patient [1].

Four types of shock are commonly defined: hypovolaemic shock (e.g. hemorrhagic shock), cardiogenic shock, distributive shock (e.g. septic shock) and obstructive shock. The mortality rate of the condition remains very high and depends on its type. It is quantified at 30% for septic shock (according to the new definition of such condition [2]) and between 35.9% and 64.7% for in-hospital mortality in patients with cardiogenic shock (depending on the stage, [3]). Usually, shock is treated in the ICU, where it is carefully monitored. As a result, large amounts of data are produced for each patient.

Machine Learning (ML) techniques can help find a list of attributes to predict the outcome of shock patients from clinical data (i.e. data routinely monitored in the ICU). However, there is a large amount of attributes that characterize each patient, ranging from base pathology information, to lab procedures or hemodynamic data and treatment, to name a few. As a result, data are likely to be very high dimensional and, therefore, feature selection (FS) methods are often required to reduce data dimensionality and its associated variability. The combination of statistics and ML may provide an appropriate framework to retrieve new knowledge from data, as well as causal relationships between different features from high-dimensional datasets [4–6].

This study analyzes the clinical database gathered within the *ShockOmics* research project. The main objective of this project was to find *'omics'* signatures of shock and it included two types of shock: septic and cardiogenic. Septic shock is the most common, and it is caused by infection. The second most frequent form of shock is the cardiogenic shock and it is a result of a heart failure. The multiscale nature of the *ShockOmics* database allows to search for attributes that are related to mortality induced by these two types of shock.

When it comes to the detection of circulatory shock, physicians and therapists largely depend on a combination of clinical, hemodynamic and biochemical signs. For a treatment to be chosen, prompt identification of shock is necessary. Appropriate treatment is based on a good understanding of the physiological mechanisms behind the condition. The management of shock is challenging and patient survival is highly dependent on the timely administration of the appropriate treatment. Normally, it requires the administration of vasoactive drugs and fluid resuscitation. This treatment is usually given to counteract the other conditions that go along with shock (e.g. hypotension/hemodynamic instability, inflammation and multiple organ failure (MOF)). The fact that they have similar symptoms makes addressing the causes of shock very difficult. As a result, decision making at the onset of the condition is not a trivial task and, therefore, it is important to design fast, reliable and interpretable tools to plan therapeutic interventions to prevent mortality, or any irreversible consequences caused by shock.

The study of the pathophysiology of shock and its management has been an active area of research in ML applications for Critical Care. For example, fuzzy decision support systems (DDS) for the management of post-surgical cardiac intensive care patients have been described in [7], while the issue of rule generation for septic shock patients has been addressed in [8] and [9], the former together with an artificial neural network (ANN). The expert system described in [10] is more closely related to the general management of sepsis and it was designed for the detection of pathogens and the prescription of antibiotics during the first hours of evolution. Ross and co-workers [11] later derived a system of ordinary differential equations together with an ANN model of inflammation and septic shock. Other studies have deployed ANNs for the study of sepsis. They include [12], who presented a clinical study examining SIRS and Multiple organ dysfunction syndrome (MODS) in the ICU after cardiac and thoracic surgery. More recently, a state-of-the-art application of a Deep Learning (DL) technique, namely Deep Reinforcement Learning (DRL) has been proposed by Raghu and co-workers [13] for the definition of continuous-space models for the treatment of sepsis, in a twist that goes beyond the more traditional development-and-use of discriminative classifiers. Support vector machine (SVM) models have also been used for the prediction of sepsis. Kim and co-workers [14] applied them to study the occurrence of sepsis in post-operative patients. Wang *et al.* [15] went a step further to build a DSS for the diagnosis of sepsis.

ML methods have also been used with varying success for the more specific problem of the prediction of mortality caused by sepsis. A diagnostic system for septic shock based on ANNs (Radial Basis Functions -RBF- and supervised Growing Neural Gas) was presented in [16]. Also in this area, Brause and colleagues [17] applied an evolutionary algorithm to an RBF network (the MEDAN Project) to obtain, over a retrospective dataset, a set of predictive attributes for assessing mortality for Abdominal Sepsis. Relevance Vector Machines (RVM), SVM variants with embedded feature selection, were used in [18], while Bayesian Networks were used in [19] and [20]. Finally, kernel methods were used in [21].

The work reported in this paper attempts to identify clinical traits that can be used as predictors of mortality in patients with cardiogenic or septic shock. Missing data imputation techniques and ML models are used, followed by different approaches to FS from different data subsets. Selected features are later used to build causal Bayesian Networks (CBN), in order to reveal potential causal relationships between features and ICU outcome.

The remaining of the paper is structured as follows. The section Materials and methods describes the *ShockOmics* dataset, as well as the FS, the ML and structure learning methods used in the reported experiments. The section Experimental Results presents the results for the FS experiments (obtained feature subsets and their classification performances) and causal discovery experiments (CBN structures). The experimental results are discussed in the section Discussion. The last section of the paper draws some conclusions about the investigated clinical problem.

Materials and methods

The ML pipeline presented in this paper is divided into two main experimental phases: FS and causal discovery.

The FS experiments deal with the problem of finding the most promising attributes (clinical traits) for the prediction of mortality related to shock. To do this, three subsets of the *ShockOmics* data were analyzed: *T1*, *T1+T2* and *Full* (described below). The goal of the experiments was to analyze different feature sets resulting from the FS procedure, and to measure their performance using ML models. The decision of starting the analyses with three different data subsets of the *ShockOmics* database was made due to the importance of obtaining a stable subset of features that behave consistently at different time-steps and that are clinically relevant. For this reason, each of the datasets targeted different features: some of them because of their low amount of missing values, and some of them because they were important at different times. There is an assumption behind the experiments, namely that the feature sets with the best performances are also the most promising for mortality prediction.

The evaluated feature sets were obtained by applying four different FS techniques to the aforementioned datasets: Univariate FS based on ANOVA F-value (*UFS*); recursive feature elimination with Support Vector Classification (*RFE*); two stage FS (*UFS+RFE*) and Random Forest (*RF*) FS. The additional fifth method was applied by averaging stability scores and aggregating features from the four previous FS techniques results. Missing values were imputed using RF imputation separately in the training and test sets. The FS techniques were performed only on the training set. The identification of the promising feature sets was based on both performance and stability scores.

Finding a subset of attributes that are predictive of mortality due to shock may not suffice in a clinical setting. The causal discovery experiments reported next allow us to go one step further and analyze the causal relationships between the data features that were selected in the previous phase of the ML pipeline. In order to reveal these causal relationships, the Fast Greedy Search (FGES) algorithm was used. The algorithm was applied twice for each of the feature subsets obtained in the FS experiments. First, the algorithm was applied only to the features, excluding the outcome, and, then, it was applied to both the features and the outcome. The expectation was to find few or no differences between features that did not involve the outcome. That would mean that the CBNs are stable enough to draw some clear conclusions. The CBNs were built using the whole dataset with the selected features imputed by RF.

For the implementation and more detailed information about the methods, please refer to the corresponding sections.

Dataset description

This study is part of the prospective observational trial ShockOmics European research project ("Multiscale approach to the identification of molecular biomarkers in acute heart failure induced by shock", Nr. 602706—ClinicalTrials.gov Identifier NCT02141607) [22] and was approved by: the Geneva Regional Research Ethics Committee (study number 14-041), the Ethical Committee of Hôpital Erasme-Université Libre de Bruxelles (study number P2014/171), and the Mútua de Terrassa Hospital Institutional Review Board (study number EO/1407). The main objective of this project was finding transcriptomics, proteomics and metabolomics signatures of shock as well as finding a robust set of features that could apply to all shock patients regardless of the initial injury. Patients admitted with both cardiogenic and septic shock to the intensive care units (ICU) of Geneva University Hospitals, Geneva, Switzerland (38-bed mixed) and Erasme Univesity Hospital, Brussels, Belgium (36-bed mixed) and Mutua de Terrassa (12-bed mixed) between October 2014 and March 2016 were screened for inclusion criteria. This study included patients diagnosed with septic or cardiogenic shock who were older than 18 years with a Sequential Organ Failure Assessment (SOFA) greater than 6, arterial lactate greater or equal than 2 mmol/l and with a documented source of infection at admission for patients with septic shock. Patients with a high risk of death within the first 24h after admission, systemic immunosuppresion, haematological diseases, metastasic cancer, pre-existing dialysis, decompensated cirrhosis or who had received more than 4 units of red blood cells or any fresh frozen plasma before screening were excluded from the study [22].

The complete database integrates blood samples and hemodynamic recordings from septic shock and cardiogenic shock patients, and from septic patients, obtained in the ICU [22]. Clinical data and blood samples were collected for analysis at: 1) T1 $<$ 16 h from T0; 2) T2 = 48 h after T0; 3) T3 = day 7 or before discharge or before discontinuation of therapy in case of fatal outcome [23]. Since a lot of patients either died or were discharged during the study, the dataset has a lot of missing values at later time-steps (especially at T3). A short summary of the clinical data can be found in Table 1.

Table 1. Clinical data summary.

Parameter	Cardiogenic shock	Septic shock	p-values
Number of patients	25	50	-
Gender (Males:Females)	20:5	34:16	-
Outcome (Alive:Dead)	19:6	38:12	-
Body mass index	26.8 ± 5.93	25.22 ± 5.54	0.2591
Total days in ICU	6.6 ± 5.42	7.56 ± 6.54	0.5289
SOFA_{T1}	10.72 ± 4.27	11.50 ± 3.7	0.4165
SOFA_{T2}	8.33 ± 4.92	8.77 ± 4.3	0.699
SOFA_{T3}	5.87 ± 3.68	6.75 ± 4.7	0.5334
APACHE II_{T1}	24.12 ± 9.79	23.16 ± 7.52	0.6396
APACHE II_{T2}	18.42 ± 9.41	16.36 ± 8.45	0.3530
APACHE II_{T3}	12.8 ± 5.35	14.83 ± 5.93	0.2727
$\text{Lactate levels}_{T1}$	6.04 ± 4.3	4.84 ± 2.58	0.1431
$\text{Lactate levels}_{T2}$	1.52 ± 0.91	2.1 ± 1.52	0.1126
$\text{Lactate levels}_{T3}$	1.25 ± 0.5	1.43 ± 0.58	0.3309

Summary of patient populations with cardiogenic and septic shock. Numerical parameters of both populations (*mean* ± *std*) are compared against each other with the Welch's t-test, where the null hypothesis is that both parameters have identical mean values.

https://doi.org/10.1371/journal.pone.0199089.t001

The *ShockOmics* dataset (data are available from the ShockOmics Institutional Data Access Board for researchers who meet the criteria for access to confidential data) has 333 features and 75 observations. 32.1% of all values are missing, and the highest percentage of missing values in one feature is 97%. Fig 1 shows a graphical depiction map of missing values for each observation. The dataset has both numerical and categorical values, but the majority of the data is numerical.

The *ShockOmics* dataset was split into three datasets. The first dataset was named *Full* and it was obtained by filtering features from the original *ShockOmics* dataset. Certain features that did not make sense from the classification viewpoint were manually removed. They were either very general comments in natural language or explicitly revealed some information about the outcome of the patient. Both cases were deemed not to be reliable as features to feed the ML model. In detail, the following features were manually removed: *ID*; *Reason for admission*; *ICU admission*; *RV area/LV area* (T1, T2, T3); *Microorganisms* (three columns with the same name); *Death due to withdrawal of care*; *Mortality 28 days, 100 days* (two columns); *Hospital result*; *Total days in ICU, in Hospital* (two columns). Moreover, the features with more

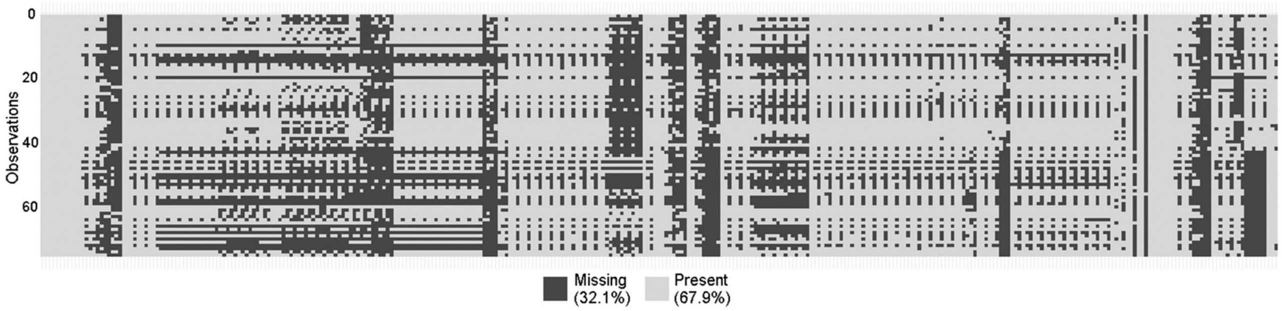

Fig 1. Graphical depiction map of missing values in *ShockOmics* dataset. The rows and the columns of the map correspond to observations and features, respectively. The black color represents missing values, the white color corresponds to present values in the *ShockOmics* dataset.

https://doi.org/10.1371/journal.pone.0199089.g001

Table 2. Initial feature set description.

Name of the feature	Observations	Type
Which type of shock	75	categorical (4 val)
Lactate levels (mmol/L)$_{T1}$	**72**	numerical (cont)
Lactate levels (mmol/L)$_{T2}$	**64**	numerical (cont)
Lactate levels (mmol/L)$_{T3}$	**38**	numerical (cont)
Mean arterial pressure (mmHg)$_{T1}$	75	numerical (cont)
Mean arterial pressure (mmHg)$_{T2}$	**71**	numerical (cont)
Mean arterial pressure (mmHg)$_{T3}$	**44**	numerical (cont)
SOFA$_{T1}$	75	numerical (cont)
SOFA$_{T2}$	**71**	numerical (cont)
SOFA$_{T3}$	**43**	numerical (cont)
APACHE II$_{T1}$	75	numerical (cont)
APACHE II$_{T2}$	71	numerical (cont)
APACHE II$_{T3}$	44	numerical (cont)
Result in ICU	75	categorical (2 val)

The columns of the table correspond to the name of the feature, the number of available observations and the type of the feature. *Initial feature set* (IFS) consists of 13 features and the target (*Result in ICU*). Features with missing observations are highlighted in bold in the *Observations* column.

https://doi.org/10.1371/journal.pone.0199089.t002

than 75% of missing data were also removed. The resulting dataset had 286 features with one target feature.

A research hypothesis is that the closer in time the feature to the final outcome, the better the prediction of mortality. To test this hypothesis, two further datasets were created, namely **T1+T2** and **T1**. Both were obtained from the *Full* dataset by filtering features at certain time-steps. From the *T1+T2* dataset all T3 features were removed, resulting in 206 features with a target feature. From the *T1* dataset both T2 and T3 time-step features were left out, so the resulting dataset included 114 features and a target feature.

A fourth feature set was built with features that are commonly assumed to be associated with the mortality of patients with shock, according to current practice. This feature set is referred to as *initial feature set* (*IFS*). Table 2 shows features that were included in the IFS set.

Feature selection and classification

In order to identify promising feature sets, five FS techniques were applied to the first three data sets described in the previous section. The resulting feature subsets were used to train ML models and five performance measures were recorded. The first four measures are defined through the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

- **Accuracy**: $\frac{TP+TN}{TP+TN+FP+FN}$;
- **Matthews correlation coefficient** ***(MCC)***: $\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$;
- **Sensitivity**: $\frac{TP}{(TP+FN)}$;
- **Specificity**: $\frac{TN}{(TN+FP)}$;
- **Area Under the Receiver Operating Characteristic Curve** ***(AUC)***. The average AUC for both labels, weighted by the number of true instances for both labels.

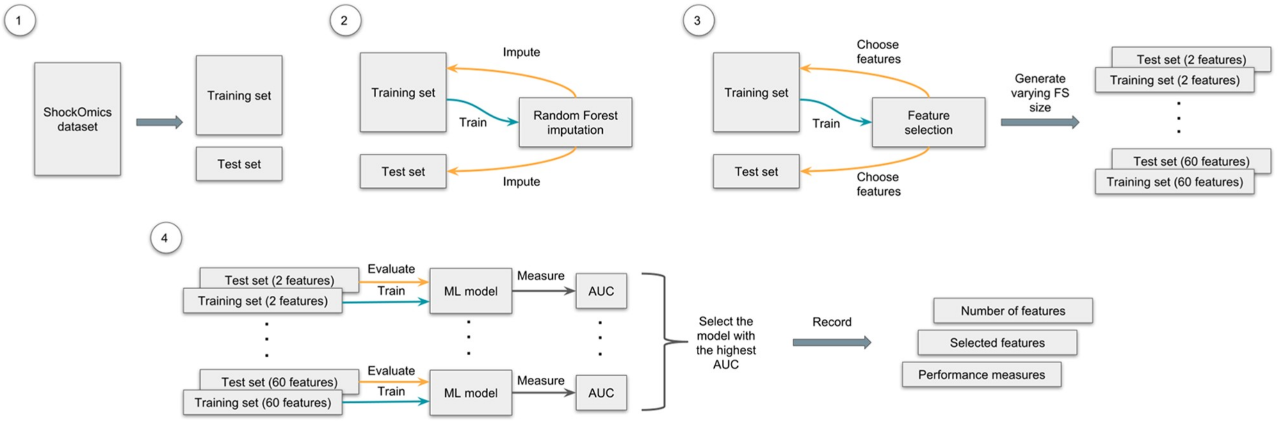

Fig 2. Graphical representation of the experimental pipeline. The process of FS and classification consists of the following steps: 1) create 100 random splits of the dataset: 75% for training and 25% for testing; 2) for each split impute both sets separately, using the imputed training set for a test set imputation; 3) after that, use the FS technique on the imputed training set, varying the size of the selected feature set (from a minimum of 2 to a maximum of 60), then choose the selected features in both sets, creating new pairs of training and test sets; 4) finally, using sets with different amounts of features, train and evaluate a ML model and choose the one with the highest AUC; record which and how many features were used for training this model, and its performance measures. Repeat these steps for all 100 random splits for each dataset and FS technique.

https://doi.org/10.1371/journal.pone.0199089.g002

The experiment pipeline for the FS procedure can be described as follows. First, 100 random splits were created for each of the datasets, using 75% for training and the rest for testing. Then, for each pair of training and test sets RF imputation [24] was used. Both sets were imputed separately one after another, using an imputed training set for a test set imputation. After that, the FS methods were applied, varying the size of the desired feature subset from a minimum of 2 to a maximum of 60. These FS models were subsequently trained on 75% of data, where only selected features were left in both training and test sets. The feature set that achieved the highest AUC was chosen as the best one for the split. The same procedure was applied to the rest of the 99 pairs of training and test sets. Once all feature sets and their performance measures were recorded, the mean and the standard deviation were calculated for each FS method and dataset. The complete experimental pipeline is graphically described in Fig 2.

Now that the best feature sets in all pairs of splits were selected, the stability scores were calculated as frequencies: the individual feature occurrences in all chosen sets were counted and then divided by the number of pairs (100 in the reported experiments). In the S1 Appendix, the 80 features with the highest stability scores are listed for each FS technique and dataset, but note that only the best 20 were later chosen to build CBNs.

The RF data imputation was performed in *R* (v3.4.3) [25], using *missForest* package that allows to impute continuous and/or categorical data including complex interactions and nonlinear relations.

The scikit-learn [26] library (v0.19.1) in Python (v2.7) was used for the implementation of the FS and ML methods. The default parameters of the methods were used in the experiments, if not otherwise specified. The five FS techniques used in the identification of promising attributes and the corresponding ML models for feature evaluation are described next:

- **Univariate feature selection based on ANOVA F-values (*UFS*).** Univariate feature selection [27] works by selecting the best features based on univariate statistical tests. In this case, ANOVA F-tests are used. The Gaussian Naive Bayes model [28] was used for the evaluation. The UFS was implemented using the sklearn.feature_selection.SelectKBest method that removes all but the *k* highest scoring features and sklearn.feature_selection.f_classif for the ANOVA F-tests. The Naive Bayes model was implemented through the sklearn.naive_bayes.GaussianNB method.

- **Recursive feature elimination with SVC (*RFE*)**. Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of the RFE is to select features by recursively considering smaller and smaller sets of features [29]. The RFE was implemented through the sklearn.feature_selection.RFE method. For the external estimator and, as the evaluated ML model, the SVC [30] was used with the help of the sklearn.svm.SVC method;
- **Two stage FS (*UFS* + *RFE*)**. A combination of two FS techniques from this list. In the first stage, UFS preselects the 120 best features (80 for the T1 dataset) based on the ANOVA F-value and then RFE (SVC) chooses the best features in the second stage. Here, the SVC was used as the ML model for the evaluation;
- **Random Forest (*RF*) FS**. This method [29] is based on the trained RF model [31] and its feature rankings. The sklearn.feature_selection.SelectFromModel method was used to retrieve the feature rankings and the sklearn.ensemble.RandomForestClassifier method to train the RF model for the FS and later (with only selected features) for evaluation (10 trees were used in each stage);
- **Aggregated (*Aggr.*) FS**. The method uses the averaged stability scores for FS obtained by the four previous techniques and, therefore, does not include the ML model. Instead, the performance is measured by averaging the measures obtained by the previous methods. This FS does not involve any processing of the dataset, but relies heavily on the previous FS techniques. As a result, it allows to achieve more stable features and analyze the aggregated performance of all trained ML models for the dataset.

Structure learning

The FGES algorithm was used to obtain the CBN structures that would reveal the interactions between the features. This is an optimized and parallelized version of the Greedy Equivalence Search algorithm. It heuristically searches the space of CBNs and returns the model with the highest value for the Bayesian Information Criterion (BIC) [32] it finds. It starts with an empty graph and adds new edges to increase the Bayesian score. The process stops when there are no edges that increase the score. It then starts removing edges until no single edge removal can increase the score.

The algorithm works on the assumption that the causal process generating the data is accurately modeled by a CBN. Each node in this CBN is a a linear function of its parents, plus a finite additive Gaussian noise term. Each observation in the data is assumed to be independent and obtained by randomly sampling all the variables from the joint distribution. Given all these assumptions, the FGES procedure outputs the CBN structure that contains:

1. an arc X → Y, if and only if X causes Y;
2. an edge (-), if and only if either X causes Y or Y causes X;
3. no edge between X and Y, if and only if X and Y have no direct causal relationship between them.

The Tetrad software (v6.4.0) was used in the implementation of FGES. The algorithm was applied twice for each of the feature sets obtained during the feature selection phase. First, the algorithm was applied only to the features, excluding the outcome, while the second time it was instead applied to both features and the outcome. The expectation was to have few or no differences between features that did not involve the outcome. That would mean that the

CBNs are stable enough to draw clear conclusions. The algorithm used the default parameters and seven maximum discrete categories.

Experimental results

Feature selection experiments

The FS experiments are divided into three groups. Each group corresponds to its own dataset: *Full*, *T1+T2* and *T1*. All five FS techniques were applied to each dataset. The results were used to train the ML model, and the performance of each feature set was evaluated for each model. In this section, the best models from all FS experiments and their accuracy, MCC, sensitivity, specificity and AUC are presented. For the details on individual features and their stability scores please refer to the S1 Appendix. Considering the 80 features with the highest stability scores for each FS technique and dataset, one may appreciate how the selected features vary across the 100 random splits.

Table 3 shows the comparison of different FS techniques and the general performance of the best features they chose: the size of the feature set, the dataset and the FS method used. The FS techniques that were applied to the same dataset are grouped together. The additional IFS dataset is used as a baseline to compare the rest of feature sets performances (the Gaussian Naive Bayes was chosen as the ML model). A naive *majority classifier* that always predicts the most frequent label in the training set, was also added for baseline comparison.

As it can be seen from Table 3, the FS models outperform manually chosen features from the IFS. Moreover, the IFS model is only slightly better than the naive *majority classifier*. The data is unbalanced and that is the reason why the accuracy and the specificity are the only measures for which the *majority classifier* is better. As for the rest of the models, they show consistently better results than the baseline. The only exception is the specificity measure, which means that there were no models that identified all negative examples correctly across all 100 random splits, and the accuracy for the *RFE* and *UFS+RFE* techniques with the *Full* dataset. The *RF* models produced consistently good results in all datasets, and were considered best for every dataset. The best result of all experiments were shown by one of these models for the *T1 +T2* dataset. Based on the *Aggr*. performance, the general performance of the models declined with the additional time-steps. This can be explained by the lower performance of the *RFE* and *UFS+RFE* models for the *T1+T2* dataset, compared to the *T1*, and by the fact that the *Full* models performed the worst across all three datasets.

For the causal discovery experiments, it was decided to choose the 20 features with the highest stability scores for the models whose results are shown in Table 4. The *RF* models showed the best performance and therefore their features required more thorough investigation. All feature sets selected for the causal discovery experiments and their individual features can be found in Box 1 below.

Within square brackets: the subset of data that was used to obtain the feature set and the applied FS technique.

Causal discovery

For each of the feature sets there are two corresponding CBNs. The first one (the CBN ‘a’ was built without the target feature (*Result in ICU*), the second one (the CBN ‘b’) included all features and the target. This section contains the CBNs and their descriptions for the *(T1, RF)* and *(T1+T2, RF)* feature sets. For the rest of CBNs, please refer to the S1 Appendix.

Fig 3 shows the CBNs for the *(T1, RF)* feature set. As it can be seen, two nodes, namely *E wave* and *E/e* at T1 are detached from the rest of the network in both CBNs. In CBN ‘a’, all edges are undirected, but the addition of the target feature, *Result in ICU*, changes most of

Table 3. The performance comparison of the feature sets.

# Features	Data	Method	Accuracy	MCC	Sensitivity	Specificity	AUC
-	-	Majority class	0.763 ± 0.09 (0.0)	0.0 ± 0.0 (0.0)	0.0 ± 0.0 (0.0)	1.0 ± 0.0 (0.0)	0.5 ± 0.0 (0.0)
11	IFS	-	0.644 ± 0.124 (0.0)	0.028 ± 0.255 (0.0)	0.269 ± 0.252 (0.0)	0.767 ± 0.14 (0.0)	0.518 ± 0.139 (0.0)
13.5 ± 13.6	T1	UFS	0.828 ± 0.083 (0.0)	0.573 ± 0.183 (0.0)	0.731 ± 0.217 (0.2496)	0.873 ± 0.095 (0.0)	0.802 ± 0.108 (0.0037)
16.5 ± 13.2	T1	RFE	0.823 ± 0.079 (0.0)	0.549 ± 0.19 (0.0)	0.702 ± 0.23 (0.0396)	0.876 ± 0.089 (0.0001)	0.789 ± 0.114 (0.0003)
14.5 ± 13.0	T1	UFS+RFE	0.814 ± 0.077 (0.0)	0.52 ± 0.181 (0.0)	0.656 ± 0.226 (0.0004)	0.876 ± 0.092 (0.0001)	0.766 ± 0.106 (0.0)
18.1 ± 10.8	T1	RF	**0.87 ± 0.07** (0.4397)	**0.668 ± 0.152** (0.268)	**0.755 ± 0.198** (0.7221)	**0.922 ± 0.074** (0.7848)	**0.839 ± 0.092** (0.6576)
15.7 ± 12.8	T1	Aggr.	0.834 ± 0.08 (0.0001)	0.577 ± 0.186 (0.0)	0.711 ± 0.221 (0.0709)	0.887 ± 0.09 (0.002)	0.799 ± 0.109 (0.0021)
10.8 ± 14.2	T1+T2	UFS	0.839 ± 0.085 (0.0008)	0.584 ± 0.207 (0.0001)	0.715 ± 0.229 (0.101)	0.889 ± 0.091 (0.0035)	0.802 ± 0.116 (0.0053)
17.5 ± 15.8	T1+T2	RFE	0.807 ± 0.089 (0.0)	0.503 ± 0.205 (0.0)	0.637 ± 0.265 (0.0002)	0.874 ± 0.101 (0.0001)	0.755 ± 0.122 (0.0)
12.9 ± 10.7	T1+T2	UFS+RFE	0.796 ± 0.074 (0.0)	0.465 ± 0.175 (0.0)	0.596 ± 0.251 (0.0)	0.873 ± 0.097 (0.0001)	0.734 ± 0.109 (0.0)
21.2 ± 14.4	T1+T2	RF	**0.878 ± 0.076** (1.0)	**0.693 ± 0.166** (1.0)	**0.765 ± 0.199** (1.0)	**0.925 ± 0.081** (1.0)	**0.845 ± 0.099** (1.0)
15.6 ± 13.9	T1+T2	Aggr.	0.83 ± 0.087 (0.0)	0.561 ± 0.208 (0.0)	0.678 ± 0.246 (0.0065)	0.89 ± 0.095 (0.0056)	0.784 ± 0.12 (0.0001)
17.9 ± 17.2	Full	UFS	0.754 ± 0.105 (0.0)	0.31 ± 0.23 (0.0)	0.431 ± 0.256 (0.0)	0.861 ± 0.129 (0.0)	0.646 ± 0.119 (0.0)
19.7 ± 16.5	Full	RFE	0.758 ± 0.088 (0.0)	0.334 ± 0.213 (0.0)	0.483 ± 0.278 (0.0)	0.855 ± 0.114 (0.0)	0.669 ± 0.118 (0.0)
17.4 ± 15.2	Full	UFS+RFE	0.743 ± 0.09 (0.0)	0.311 ± 0.21 (0.0)	0.477 ± 0.251 (0.0)	0.839 ± 0.123 (0.0)	0.658 ± 0.111 (0.0)
21.9 ± 15.5	Full	RF	**0.819 ± 0.082** (0.0)	**0.53 ± 0.176** (0.0)	**0.614 ± 0.244** (0.0)	**0.902 ± 0.087** (0.0544)	**0.758 ± 0.111** (0.0)
19.2 ± 16.2	Full	Aggr.	0.769 ± 0.096 (0.0)	0.371 ± 0.228 (0.0)	0.501 ± 0.267 (0.0)	0.864 ± 0.117 (0.0)	0.683 ± 0.123 (0.0)

Best feature sets obtained in the FS experiments and their general performance across 100 random data splits. The columns correspond to the number of features, the data that were used to obtain the feature set, the FS method and five performance measures ($mean \pm std(p - value)$). Welch's t-test was used to obtain p-values for the null hypothesis that two performance measures had identical values. Each measure was tested against the same measures of the (T1+T2, RF)) model. The *IFS* dataset was used as a baseline for comparison. The best results for all measures in each dataset group are highlighted in bold.

https://doi.org/10.1371/journal.pone.0199089.t003

them to directed. Moreover, adding this feature replaces the edge between *Lactate levels* and *Norepinephrine* at T1 to the edge between *K Ur* and *Na Ur* at T1, connecting these two nodes to the CBN. According to the CBN 'b', *Result in ICU* can be potentially influenced by *Norepinephrine*, *Respiratory rate*, *Platelet count* and *SOFA* at T1.

Fig 4 shows the CBNs for the *(T1+T2, RF)* feature set. In both CBNs, no edges are connected to *K Ur* at T1. This time, the addtion of the outcome attaches *E wave* at T1 and *Neutro abs count* at T2 to the rest of the network. There are multiple changes across the CBN that involve other features as well. For example, edges between *Glasgow Coma Scale* at T1 and *Lactate levels* at T2, and between *Lactate levels* at T2 and *PT* at T2 disappear. Moreover, for all edges the causal directions become clear, and new edges emerge. The CBN 'b' shows that

Table 4. The performance of the feature sets selected for the causal discovery experiments.

# Features	Data	Method	Accuracy	MCC	Sensitivity	Specificity	AUC
18.1 ± 10.8	T1	RF	0.87 ± 0.07 (0.4397)	0.668 ± 0.152 (0.268)	0.755 ± 0.198 (0.7221)	0.922 ± 0.074 (0.7848)	0.839 ± 0.092 (0.6576)
15.7 ± 12.8	T1	Aggr.	0.834 ± 0.08 (0.0001)	0.577 ± 0.186 (0.0)	0.711 ± 0.221 (0.0709)	0.887 ± 0.09 (0.002)	0.799 ± 0.109 (0.0021)
21.2 ± 14.4	T1+T2	RF	**0.878 ± 0.076** (1.0)	**0.693 ± 0.166** (1.0)	**0.765 ± 0.199** (1.0)	**0.925 ± 0.081** (1.0)	**0.845 ± 0.099** (1.0)
15.6 ± 13.9	T1+T2	Aggr.	0.83 ± 0.087 (0.0)	0.561 ± 0.208 (0.0)	0.678 ± 0.246 (0.0065)	0.89 ± 0.095 (0.0056)	0.784 ± 0.12 (0.0001)
21.9 ± 15.5	Full	RF	0.819 ± 0.082 (0.0)	0.53 ± 0.176 (0.0)	0.614 ± 0.244 (0.0)	0.902 ± 0.087 (0.0544)	0.758 ± 0.111 (0.0)
19.2 ± 16.2	Full	Aggr.	0.769 ± 0.096 (0.0)	0.371 ± 0.228 (0.0)	0.501 ± 0.267 (0.0)	0.864 ± 0.117 (0.0)	0.683 ± 0.123 (0.0)

The columns of the table correspond to the number of features, the data that was used to obtain the feature set, the FS method used and five performance measures (*mean* ± *std*(*p* value)). Each measure was tested against the same measures of the *(T1+T2, RF)* model. The best results in the performance measures columns are highlighted in bold.

https://doi.org/10.1371/journal.pone.0199089.t004

Result in ICU is potentially influenced by *E wave* at T1, *Norepinephrine* at T2, *Neutro abs count* at T2 and *SOFA* at T2.

In most of CBNs pairs, the presence of the target value did not imply much change. The majority of edges of the *Full* models were undirected and the parts of the network were detached from each other. Adding *Result in ICU* does not connect these separate parts. This contrasts with the *T1* and *T1+T2* models, that have more or less connected CBNs. As for the *T1* and `T1+T2` CBNs, they have common patterns that are noticeable throughout the experiments. More importantly, the two best models described in this section have some common features that are close to the outcome in both CBNs. These similarities are discussed in the next section.

Discussion

Analyzing the results of the FS experiments, one may notice that there are a lot of reoccurring features with high stability scores for each of the datasets. These features can be clearly seen from the *Aggr.* tables in page 19. For *T1* the most common features are *APACHE II*, *SOFA*, *respiratory rate*, *Glasgow coma scale* and *K Ur*. All these features except *Glasgow coma scale* scored high with the model that showed the best results for this dataset. Moreover, some of them were also present in the results for *T1+T2*. Interestingly, there are many similar features that scored high in both *T1* and *T1+T2*, which is a clear indication of their importance for the mortality prediction task.

Regarding the *T1+T2* dataset, the most stable features are *APACHE II* at T1, *SOFA* at T2, *tricuspid regurgitation maximal velocity* at T1, *Glasgow Coma Scale* at T1 and *neutrophils absolute count* at T2. It is especially interesting that some attributes at T1 occurred more frequently than their T2 counterparts. This can be explained by the fact that features at later time-steps present more noise and missing features. However, the *(T1+T2, RF)* showed the best predictive power in our experiments, which means that at least some T2 features improved the results obtained.

In contrast, the worst performance was that obtained for the *Full* dataset. Its most stable features include *A wave* at T2, *systolic blood pressure* at T2, *heart rate* at T1, *respiratory rate* at T1 and *weight*. These features scored the highest, but their stability scores are very low. Moreover,

Box 1. List of best feature selections

[T1, RF]: APACHE II T1, SOFA T1, Respiratory rate T1, Urine Output (mL/day) T1, E wave (cm/s) T1, Fluid Balance (ml) T1, K Ur T1, Lactate levels (mmol/L) T1, Tidal volume (VT) T1, Pulmonary artery systolic pressure (TR jet by CW + CVP) (mmHg) T1, * Norepinephrine (mg/kg/min) T1, PCT Value (mg/mL) T1, Heart rate (bpm) T1, E/e' T1, PEEP T1, Pplat T1, Platelet count T1, Na Ur T1, pH T1, PaO2/FiO2 T1;

[T1, Aggr.]: APACHE II T1, SOFA T1, Respiratory rate T1, Glasgow Coma Scale T1, K Ur T1, Platelet count T1, Tidal volume (VT) T1, Tricuspid regurgitation maximal velocity (by CW) (cm/s) T1, HCO3 (mmol/L) T1, LA dilatation by eyeballing T1, Pulmonary artery systolic pressure (TR jet by CW + CVP) (mmHg) T1, Base Excess (mmol/L) T1, E wave (cm/s) T1, PCT Value (mg/mL) T1, Respiratory rate (rpm) T1, Pplat T1, Tricuspid annular tissular doppler S wave (DTI) (cm/s) T1, * Norepinephrine (mg/kg/min) T1, * Dobutamine (mg/kg/min) T1, Lactate levels (mmol/L) T1;

[T1+T2, RF]: Lactate levels (mmol/L) T2, SOFA T2, Tidal volume (VT) T2, APACHE II T1, Urine Output (mL/day) T1, Urine Output (mL/day) T2, APACHE II T2, Respiratory rate T1, K Ur T2, SOFA T1, * Norepinephrine (mg/kg/min) T2, Neutro abs count T2, E wave (cm/s) T1, Na Ur T2, pH T2, K Ur T1, Glasgow Coma Scale T2, Tricuspid regurgitation maximal velocity (by CW) (cm/s) T1, PT T2, Pulmonary artery systolic pressure (TR jet by CW + CVP) (mmHg) T1;

[T1+T2, Aggr]: APACHE II T1, SOFA T2, Tricuspid regurgitation maximal velocity (by CW) (cm/s) T1, Glasgow Coma Scale T1, Neutro abs count T2, Glasgow Coma Scale T2, Platelet count T1, K Ur T1, APACHE II T2, SOFA T1, Respiratory rate T1, HCO3 (mmol/L) T1, Lactate levels (mmol/L) T2, E wave (cm/s) T1, Tidal volume (VT) T1, Respiratory rate (rpm) T1, K Ur T2, Na Ur T2, PT T2, LA dilatation by eyeballing T1;

[Full, RF]: Neutro abs count T1, Platelets (10 3/mm 3) T3, E/e' T2, Lateral e' (cm/s) T2, Creat Ur T2, Urine Output (mL/day) T2, Lateral e' (cm/s) T1, Respiratory rate (rpm) T1, Lympho abs count T1, Mean arterial pressure (mmHg) T1, Diastolic Blood Pressure (mmHg) T2, E wave deceleration time (ms) T2, Tricuspid annular tissular doppler S wave (DTI) (cm/s) T1, Sat O2/FiO2 T2, Platelets (10 3/mm 3) T1, FiO2 T3, Tidal volume (VT) T1, RBC count T3, Fluid Balance (ml) T2, Platelet count T2;

[Full, Aggr]: A wave (cm/s) T2, Systolic blood pressure (mmHg) T2, Heart rate (bpm) T1, Respiratory rate (rpm) T1, Weight (kg), Height (cm), A wave (cm/s) T1, PT T1, E wave deceleration time (ms) T3, Hematocrit (%) T2, Platelet count T3, Creat Ur T1, E wave deceleration time (ms) T1, PaCO2 (mmHg) T1, K Ur T1, E wave (cm/s) T2, aPTT T1, Mean arterial pressure (mmHg) T1, Base Excess (mmol/L) T1, PaO2 (mmHg) T2.

the difference between all scores of the *Full* dataset is very small. This means that during the FS procedures across 100 random experiments, very different features were selected making them unreliable for their low stability. This supports the hypothesis that adding more time-steps (i.e., later times) increases noise, and that the current amount of data is not sufficient enough to bypass this problem.

The performance of features from different datasets and models were further compared to the model that showed the best results, which was the *(T1+T2, RF)* model. More specifically, the performance measures of *(T1+T2, RF)* were tested against the measures of other models,

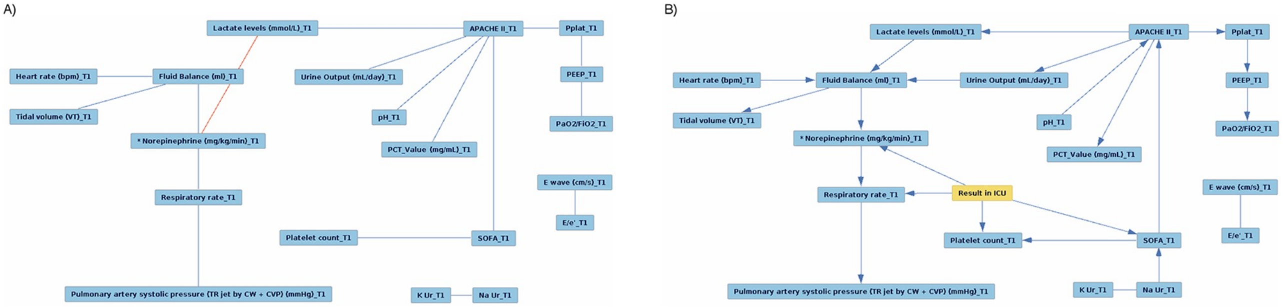

Fig 3. The *(T1, RF)* CBN. The CBN for the *(T1, RF)* feature set, obtained: a) without the target feature; b) with the target feature.

https://doi.org/10.1371/journal.pone.0199089.g003

using the Welch's t-test. The null hypothesis was that each pair of performance measures has identical values. Setting the threshold for the p-value to 0.05 shows that there is a significant chance that the *(T1, RF)* and *(T1+T2, RF)* models perform similarly according to all five performance measures. For this reason, these models were singled out in the causal discovery experiments.

The CBN structures revealed connections between the features and the outcome. The assumption behind the evaluation of the most important clinical attributes for outcome prediction was that such features are the closest to the outcome in a graph. The direction of the edge from feature to the outcome is also a good indicator that a certain feature is important.

According to the *RF* and *Aggr*. CBNs for the *T1* dataset, *Result in ICU* is influenced by *norepinephrine* and *SOFA* at T1. Both models presented direct connections between these two features and the outcome. The additional possible candidates, supported by the *(T1, RF)*, are *respiratory rate* and *platelet count* at T1. *E wave*, *dobutamine* and *tricuspod regurgitation maximal velocity* at T1 were considered important by the *(T1, Aggr.)* pair. As for the *T1+T2* dataset, its *RF* and *Aggr*. models show the significance of *neutrophils absolute count* at T2. Additional features that are directly connected to *Result in ICU* in the *RF* CBN include *norepinephrine* at T2, *E wave* at T1 and *SOFA* at T2. The *Aggr*. model adds *APACHE II* at T2 to the list. The *Full* CBNs show that the selected features do not create any coherent structure and, hence, cannot be used to draw any conclusions. This can be explained by the low stability scores and the lack of variability among them in the previous experiments.

Some parallels can be seen between the features that achieved high *Aggr*. stability scores and those that were considered important in the causal discovery experiments. These features were ranked very high by the FS techniques and had the direct connection to the outcome. These features include *SOFA* at T1, *respiratory rate* at T1, *tricuspid regurgitation maximal velocity* at

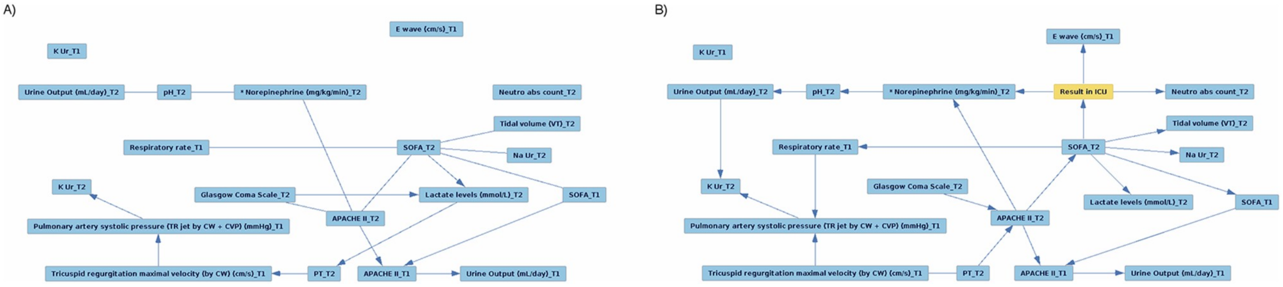

Fig 4. The *(T1+T2, RF)* CBN. The CBN for the *(T1+T2, RF)* feature set obtained: a) without the target feature; b) with the target feature.

https://doi.org/10.1371/journal.pone.0199089.g004

Box 2. List of promising attributes

High stability scores features: APACHE II T1, SOFA T1, Respiratory rate T1, Glasgow Coma Scale T1, K Ur T1, SOFA T2, Tricuspid regurgitation maximal velocity T1, Neutro abs count T2, Glasgow Coma Scale T2.

Causal discovery features: Norepinephrine T1, SOFA T1, Respiratory rate T1, Platelet count T1, E wave at T1, Dobutamine T1, Tricuspod regurgitation maximal velocity T1, Neutro abs count T2, Norepinephrine T2, SOFA T2, APACHE II T2.

T1, *neutrophils absolute count* at T2 and *SOFA* at T2. Additional features that were in the top 20 of the *Aggr*. stability scores tables but were ranked lower are *norepinephrine* at T1, *platelet count* at T1, *E wave* at T1 and *dobutamine* at T1. The full list of features that were considered valuable for the mortality prediction task can be found in Box 2. Since the *Full* features had very low stability scores, only the first few T1 and T1+T2 features from *Aggr*. tables were considered important.

List of attributes that were considered promising for the outcome prediction in patients with shock.

These findings agree with other studies that support the use of certain attributes for mortality prediction. For instance, the SOFA score has previously been shown to have a significant prognostic value for in-hospital mortality prediction [33] [34]. The APACHE II score is often used for outcome prediction too [35–37]. And in the study by Jansen *et al.* [38], targeting a decrease of at least 20% in the blood lactate level (a common feature for the best two models) over a 2-hour period seemed to be associated with reduced in-hospital mortality in patients with shock. However, some of the attributes that are also related to mortality in patients with shock, like mean blood pressure [39], mean arterial pressure [40], or low LVEF [41] did not show significant predictive value in this study.

Finally, it should be noted that some features seem more important at certain time-steps. For example, the difference in stability scores shows that *neutrophils absolute count* at T2, *tricuspid regurgitation maximal velocity* at T1 and *dobutamine* at T2 were selected far more often than their counterparts. This might be due to the specifics of imputation or due to the lack of data. In any case, this hypothesis requires a more thorough investigation, that is beyond the scope of this work.

Conclusion

In this paper, we have analyzed and experimentally compared several models for the prediction of mortality in patients with cardiogenic or septic shock, identifying those clinical traits that are more relevant for such prediction in the ICU at different time-points. These models are the result of the application of different FS techniques over the available clinical data. In particular, we were interested in obtaining an actionable classifier for the acute phase of shock, which is the most critical to implement an appropriate therapeutic response and, therefore, has the greater impact on the prognosis of patients. For the acute phase (i.e. the first 24 hours of evolution of shock), we have obtained a classifier that leverages upon well established attributes for septic shock such as the SOFA or APACHE II scores and treatment with vasopressors (for example, norepinephrine), which are part of the cardiovascular SOFA score. This classifier for the acute phase of shock also resorts to other valuable attributes such as those obtained through echocardiography during cardiogenic shock (e.g. E wave velocity, early mitral inflow velocity and mitral annular early diastolic velocity (E/e')). Other important attributes are related to respiratory function (such as respiratory rate, tidal volume and ventilator pressures), fluid balance, pulmonary artery pressure, renal function (such as urine output and urea levels), lactate levels, acidosis and C-reactive protein levels. These significant attributes are very closely related to the new official definition of shock and its corresponding management guidelines, in particular those related to organ dysfunction and respiratory function. However, our causal discovery experiments showed that treatment with pressors, respiratory rate, platelet count and the SOFA score presented the highest dependence with ICU outcome. The best classifier for the acute phase yielded an accuracy of 0.870, a sensitivity of 0.755 and a specificity of 0.922. The accuracy results were not statistically different from those obtained with the best classifier at T1 and T2.

The main conclusion of this study is that it is possible to predict the risk of death in the acute phase of septic and cardiogenic shock with quite acceptable results by taking into consideration the attributes routinely measured through echocardiography in the ICU. Use of this data for assessing the prognosis of patients is considered valuable for the clinical management of patients with shock in the ICU at ICU admission or during the first day of evolution.

The results presented here open up a new line of research for the study of the pathophysiology of shock when combined with other data related to organ dysfunction beyond the commonly used scales such as the SOFA score. In the next step of this research, we plan to add other types of data that are related to the patient's response to shock and assess to what extend the expanded models may yield improved results. This new layer of data will include transcriptomics, proteomics and metabolomics features as detailed in the scope of the ShockOmics European project.

Supporting information

S1 Appendix. Complementary material for the FS and causal discovery experiments. The resulting features from applying multiple FS techniques to the three available datasets, their stability scores, additional CBNs and a full list of ShockOmics attributes.
(PDF)

Acknowledgments

This research was supported by the European FP7 project Shockomics (Nr. 602706), Multiscale approach to the identification of molecular biomarkers in acute heart failure induced by shock, and Spanish research project TIN2016-79576-R.

Author Contributions

Conceptualization: Vicent Ribas Ripoll, Alfredo Vellido.

Data curation: Vicent Ribas Ripoll.

Formal analysis: Alexander Aushev.

Investigation: Alexander Aushev.

Methodology: Alexander Aushev, Vicent Ribas Ripoll, Alfredo Vellido.

Project administration: Giuseppe Baselli.

Software: Alexander Aushev.

Supervision: Vicent Ribas Ripoll, Alfredo Vellido.

Validation: Alexander Aushev.

Visualization: Alexander Aushev.

Writing – original draft: Alexander Aushev, Vicent Ribas Ripoll.

Writing – review & editing: Alfredo Vellido, Federico Aletti, Bernardo Bollen Pinto, Antoine Herpain, Emiel Hendrik Post, Eduardo Romay Medina, Ricard Ferrer, Karim Bendjelid.

References

1. Vincent JL, De Backer D. Circulatory shock. New England Journal of Medicine. 2013; 369(18):1726–1734. https://doi.org/10.1056/NEJMra1208943 PMID: 24171518
2. Sterling SA, Puskarich MA, Glass AF, Guirgis F, Jones AE. The impact of the Sepsis-3 septic shock definition on previously defined septic shock patients. Critical care medicine. 2017; 45(9):1436–1442. https://doi.org/10.1097/CCM.0000000000002512 PMID: 28542029
3. Nguyen HL, Yarzebski J, Lessard D, Gore JM, McManus DD, Goldberg RJ. Ten-Year (2001–2011) Trends in the Incidence Rates and Short-Term Outcomes of Early Versus Late Onset Cardiogenic Shock After Hospitalization for Acute Myocardial Infarction. Journal of the American Heart Association. 2017; 6(6):e005566. https://doi.org/10.1161/JAHA.117.005566 PMID: 28592462
4. McNally R, Mair P, Mugno B, Riemann B. Co-morbid obsessive–compulsive disorder and depression: A Bayesian network approach. Psychological medicine. 2017; 47(7):1204–1214. https://doi.org/10.1017/S0033291716003287 PMID: 28052778
5. Luo Y, El Naqa I, McShan DL, Ray D, Lohse I, Matuszak MM, et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis. Radiotherapy and Oncology. 2017; 123(1):85–92. https://doi.org/10.1016/j.radonc.2017.02.004 PMID: 28237401
6. Cooper G, Cai C, Lu X. Tumor-specific Causal Inference (TCI): A Bayesian Method for Identifying Causative Genome Alterations within Individual Tumors. bioRxiv. 2018; p. 225631.
7. Denai M, Mahfouf M, Ross J. A fuzzy decision support system for therapy administration in cardiovascular intensive care patients. In: Proceedings of the FUZZ-IEEE; 2007. p.1-6.
8. Paetz H. Metric rule generation with septic shock patient data. In: Proceedings of the ICDM; 2001. p.637-8.
9. Paetz J. Intersection based generalization rules for the analysis of symbolic septic shock patient data. In: Proceedings of the ICDM; 2002. p. 673-6.
10. Duhamel A, Beuscart R, Demongeot J, Mouton Y. SES (Septicemia Expert System): knowledge validation from data analysis. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS) vol.3. 1988; p.1400-1.
11. Ross JJ, Mason DG, Paterson IG, Linkens DA, Edwards ND. Development of a knowledge-based simulator for haemodynamic support of septic shock. In: IEEE Colloquium on Simulation in Medicine (Ref. No. 1998/256). 1998; 3/1-3/4.
12. Schuh ChJ. Sepsis and septic shock analysis using neural networks. In: Proceedings of the Annual Meeting of the NAFIPS; 2007. p. 650-4.
13. Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous State-Space Models for Optimal Sepsis Treatment-a Deep Reinforcement Learning Approach. arXiv preprint. 2017; arXiv:1705.08422.
14. Kim J, Blum J, Scott, C. Temporal features and kernel methods for predicting sepsis in postoperative patients. Technical Report, University of Michigan, USA. 2010.
15. Wang SL, Wu F, Wang BH. Prediction of severe sepsis using SVM model. In: Advances in Experimental Medicine and Biology Series, vol.680(1), Springer; 2010. p. 75–81. https://doi.org/10.1007/978-1-4419-5913-3_9
16. Brause R, Hamker F, Paetz J. Septic shock diagnosis by neural networks and rule based systems. Studies in Fuzziness and Soft Computing, vol.96; 2002. p.323–56. https://doi.org/10.1007/978-3-7908-1788-1_12
17. Brause R, Hanisch E, Paetz J, Arlt B. Neural networks for sepsis prediction—the MEDAN project. Journal fur Anasthesie und Intensivbehandlung. 2004; 11:40–3.
18. Ribas V, Ruiz-Rodríguez JC, Wojdel A, Caballero-López J, Ruiz-Sanmartín A, Rello J, Vellido A. Severe sepsis mortality prediction with relevance vector machines. In: Proceedings of the 33rd Annual

International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2011. p.100-3.

19. Ribas VJ, Caballero-López J, Sáez de Tejada A, Ruiz-Rodríguez JC, Ruiz-Sanmartín A, Rello J, Vellido A. On the use of graphical models to study ICU outcome prediction in septic patients treated with statins. In: Biganzoli E, Vellido A, Ambrogi F, Tagliaferri R, editors. CIBB 2011, LNCS/LNBI vol.7548; 2012. p.98–111.

20. Morales C, Vellido A, Ribas V. Applying conditional independence maps to improve sepsis prognosis. In: Data Mining in Biomedical Informatics and Healthcare (DMBIH) Workshop. IEEE International Conference on Data Mining (ICDM), 2016.

21. Ribas V, Vellido A, Romero E, Ruiz-Rodríguez JC. Sepsis mortality prediction with quotient basis kernels. Artif. Intell. Med. 2014; 61(1):45–52. https://doi.org/10.1016/j.artmed.2014.03.004

22. Ferrario M, Aletti F, Baselli G. The EU ShockOmics Project International Workshop at ICCAI'17; 2017.

23. Aletti F, Conti C, Ferrario M, Ribas V, Pinto BB, Herpain A, et al. ShockOmics: multiscale approach to the identification of molecular biomarkers in acute heart failure induced by shock. Scandinavian journal of trauma, resuscitation and emergency medicine. 2016; 24(1):9. https://doi.org/10.1186/s13049-016-0197-4 PMID: 26822963

24. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2011; 28(1):112–118. https://doi.org/10.1093/bioinformatics/btr597 PMID: 22039212

25. R Core Team. R: A Language and Environment for Statistical Computing; 2017. Available from: https://www.R-project.org/.

26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

27. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007; 23(19):2507–2517. https://doi.org/10.1093/bioinformatics/btm344 PMID: 17720704

28. Chan TF, Golub GH, LeVeque RJ. Updating formulae and a pairwise algorithm for computing sample variances. In: COMPSTAT 1982 5th Symposium held at Toulouse 1982. Springer; 1982. p. 30–41.

29. Lin X, Wang Q, Yin P, Tang L, Tan Y, Li H, et al. A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. Metabolomics. 2011; 7(4):549–558. https://doi.org/10.1007/s11306-011-0274-7

30. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural processing letters. 1999; 9(3):293–300. https://doi.org/10.1023/A:1018628609742

31. Liaw A, Wiener M, et al. Classification and regression by randomForest. R news. 2002; 2(3):18–22.

32. Raftery AE. Bayesian model selection in social research. Sociological methodology. 1995; p. 111–163. https://doi.org/10.2307/271063

33. Rivera-Fernández R, Nap R, Vázquez-Mata G, Miranda DR. Analysis of physiologic alterations in intensive care unit patients and their relationship with mortality. Journal of critical care. 2007; 22(2):120–128. https://doi.org/10.1016/j.jcrc.2006.09.005 PMID: 17548023

34. Raith EP, Udy AA, Bailey M, McGloughlin S, MacIsaac C, Bellomo R, et al. Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. Jama. 2017; 317(3):290–300. https://doi.org/10.1001/jama.2016.20328 PMID: 28114553

35. Huang Y, Chen J, Zhong S, Yuan J. Role of APACHE II scoring system in the prediction of severity and outcome of acute intracerebral hemorrhage. International Journal of Neuroscience. 2016; 126 (11):1020–1024. https://doi.org/10.3109/00207454.2015.1099099 PMID: 26393395

36. Agarwal A, Choudhary GS, Bairwa M, Choudhary A. Apache II scoring in predicting surgical outcome in patients of perforation peritonitis. International Surgery Journal. 2017; 4(7):2321–2325. https://doi.org/10.18203/2349-2902.isj20172790

37. Lee H, Lim C, Hong H, Ju J, Jeon Y, Hwang J, et al. Efficacy of the APACHE II score at ICU discharge in predicting post-ICU mortality and ICU readmission in critically ill surgical patients. Anaesthesia & Intensive Care. 2015; 43(2).

38. Jansen TC, van Bommel J, Schoonderbeek FJ, Sleeswijk Visser SJ, van der Klooster JM, Lima AP, et al. Early lactate-guided therapy in intensive care unit patients: a multicenter, open-label, randomized controlled trial. American journal of respiratory and critical care medicine. 2010; 182(6):752–761. https://doi.org/10.1164/rccm.200912-1918OC PMID: 20463176

39. Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, et al. Developing a new definition and assessing new clinical criteria for septic shock: for the Third International Consensus

Definitions for Sepsis and Septic Shock (Sepsis-3). Jama. 2016; 315(8):775–787. https://doi.org/10.1001/jama.2016.0289 PMID: 26903336

40. Leone M, Asfar P, Radermacher P, Vincent JL, Martin C. Optimizing mean arterial pressure in septic shock: a critical reappraisal of the literature. Critical Care. 2015; 19(1):101. https://doi.org/10.1186/s13054-015-0794-z PMID: 25888071

41. Harjola VP, Lassus J, Sionis A, Køber L, Tarvasmäki T, Spinar J, et al. Clinical picture and risk prediction of short-term mortality in cardiogenic shock. European journal of heart failure. 2015; 17(5):501–509. https://doi.org/10.1002/ejhf.260 PMID: 25820680