scientific reports

OPEN

Survival prediction of patients with sepsis from age, sex, and septic episode number alone

Davide Chicco[1]✉ & Giuseppe Jurman[2]

Sepsis is a life-threatening condition caused by an exaggerated reaction of the body to an infection, that leads to organ failure or even death. Since sepsis can kill a patient even in just one hour, survival prediction is an urgent priority among the medical community: even if laboratory tests and hospital analyses can provide insightful information about the patient, in fact, they might not come in time to allow medical doctors to recognize an immediate death risk and treat it properly. In this context, machine learning can be useful to predict survival of patients within minutes, especially when applied to few medical features easily retrievable. In this study, we show that it is possible to achieve this goal by applying computational intelligence algorithms to three features of patients with sepsis, recorded at hospital admission: sex, age, and septic episode number. We applied several data mining methods to a cohort of 110,204 admissions of patients, and obtained high prediction scores both on this complete dataset (top precision-recall area under the curve PR AUC = 0.966) and on its subset related to the recent Sepsis-3 definition (top PR AUC = 0.860). Additionally, we tested our models on an external validation cohort of 137 patients, and achieved good results in this case too (top PR AUC = 0.863), confirming the generalizability of our approach. Our results can have a huge impact on clinical settings, allowing physicians to forecast the survival of patients by sex, age, and septic episode number alone.

Sepsis is a dangerous condition triggered by an immune overreaction to an infection. According to the World Health Organization (WHO) estimates, sepsis affects more than 30 million people yearly worldwide , causing approximately 6 million deaths[1], and causing more than US$24 billion healthcare related costs annually just in United States[2]. The scientific community is still investigating sepsis etiology[3], whilst its management[4–6] is troublesome due to the high disease's complexity and heterogeneity[7,8]. A further complexity factor lies in a more restrictive definition of sepsis introduced in 2016[9]; the new definition, named Sepsis-3[10], now requires the presence of additional organ dysfunctions for the condition to be labelled as sepsis. Although the usefulness of Sepsis-3 has recently been validated[11], it is still debated within the medical community[12]. Additionally, early detection is critical to managing the attack and obtaining a favorable outcome, as Sepsis can kill a patient in as little as an hour.

Prediction of survival of patients with sepsis. Medical literature is rich of general purpose articles on sepsis[13], and quest for biomarkers in clinical settings have now spanned several decades, with papers dating back to early seventies still relevant today[14]. Initially, the core of the researches focused on clinical trials aimed at identifying therapeutic factors representing potential targets for novel or repurposed drugs. The crucial change of pace occurred in the early 2000s, when broad epidemiological data begun being publicly available, yielding the appearance of large retrospective studies[2,15]. Indeed, such recent influx of data has resulted in a steady flow of medical and computer science studies in which researchers have used various data science techniques to find associations between clinical factors and sepsis outcomes, with patient survival being among the most important. Contributing to the landscape, the practitioners' community started introducing different early warning scores, such as physiological monitoring systems for detecting of acutely deteriorating patients[16]. A small group of scores quickly gained popularity in the clinical settings, thus becoming *de facto* standards for benchmarking studies: APACHE[17], SAPS[18], SOFA[19] and qSOFA score[10]. Adding to such established community shared scorxes, different formulas have been recently defined in the literature involving alternative variables: for instance, the dynamic pulse pressure and vasopressor (DPV), the delta pulse pressure (ΔPP)[20] and the sepsis hospital mortality score (SHMS)[21]. However, although early warning scores have been widely adopted, there is only limited

[1]Krembil Research Institute, Toronto, ON, Canada. [2]Fondazione Bruno Kessler, Trento, Italy. ✉email: davidechicco@davidechicco.it

evidence of their effectiveness in the improvement of patient outcome[16]. Among all statistical methods, algorithms based on multivariate (Cox) regression on clinical variables have played a key role[22,23], since back in the early years[24] to nowadays[25]. Notably, the features involved in these methods are not limited to clinical variables: in the last few years a number of teams tried alternative elements from modern omics technologies, such as metabolomics[26], SNPs genomics[27], circulating microRNA[28], blood metabolites[29] or lymphocytes apoptosis[30], often coupled with more classical biomarkers and compared with the different scores. Unfortunately, these statistical based approaches proved to be rather limited in their performances, with only a tiny fraction of studies achieving acceptable level of efficacy[31]. Indeed, Gwadry-Sridhar and colleagues[32] claimed superiority of decision trees over regression methods already in 2010.

Sepsis and machine learning.

More recently machine learning has become the major player in the predictive analysis of sepsis data, leading to a massive wave of studies targeting different aspects of the problem, from the general issue[33–51], to more specific objectives or methods. For instance, many studies have been defining, combining and validating score risks[52,53], predicting early onset[54–56], or focusing on pediatric aspects[57] or on the immediate applicability to clinical practice[58–60]. Longitudinal studies have also appeared[61–63], together with methods integrating alternative data sources such as omics[64] and others. In the end of the 2010s, the computational intelligence revolution entered the playground too, and deep learning approaches flooded the specialized journals[65–73], also considering the interpretability issue[74,75]. Fleuren et al.[76] published comprehensive review of the different aspects. As mentioned earlier, many of these studies have been possible thanks to the public availability of curated clinical datasets related to sepsis. Among these datasets, we point out the Surviving Sepsis Campaign initiative[77,78] (albeit not fully publicly released), the Medical Information Mart for Intensive Care database (MIMIC-III)[79] and the electronic Intensive Care Unit (eICU) Collaborative Research Database[80], which stand our for their completeness and integrity. Additionally, it is worth mentioning some notable studies aimed at identifying a restricted number of sepsis survival predicting features[81,82]: for instance, six predictors by Mao and colleagues[83,84], five main predictive features by Shukeri et al.[85], and three blood biomarkers by Dolin and coauthors[86].

This study.

In the present manuscript, we take a similar approach: our driving goal is the prediction of the binary survival in a large cohort of Norwegian patients originally introduced and made public by Knoop and colleagues[87]. In addition to this prognostic task, as a distinguishing feature we also aim at proving that a minimal set of predictors can adequately predict the survival status. To further confirm the validity of our approach, we show that our approach can also be applied to an external South Korean dataset having the same clinical features, that we use as validation cohort. As a major outcome of such quest and improving over the published literature, we discovered that a single clinical factor, namely the progressive hospitalization episode, coupled with the two basic personal elements age and sex, can effectively predict the survival of the patients. Notably, we carried out the analysis both on the whole cohort, originally called *primary cohort*, corresponding to the admissions of the patients affected by sepsis potential preconditions (ante Sepsis-3 definition), and on a subset of the data including only the patients' admissions defined by the novel Sepsis-3 definition, called *study cohort*. We then repeated the same analysis entirely on the validation cohort, and finally trained our models on the primary and study cohorts to apply them afterwards to the validation cohort. For the first time, we show that is possible to apply machine learning to sex, age, and septic episode number collected from admission clinical records to predict the survival of the patients who had sepsis. Our very small set of detected predictors represent a sensible compromise between accuracy and simplicity of the model, requiring few resources as collected data. This balance is critical when considering the translation to clinical practice, which especially for sepsis management is rarely successful[58] and not easily integrated with clinicians' activities[88]. Although a number of digital handling proposals have appeared in the literature[89], the impact of sepsis of Food and Drugs Administration (FDA) approved Software as Medical Devices (SaMD)[90], for example, is yet far from being widespread, with perhaps the Sepsis Prediction and Optimization of Therapy system (SPOT)[91] as the most famous example. This given, having a simple albeit accurate predictive test on patient survival as presented here is a promising initial step towards the development of a machine learning-based tool supporting clinicians in everyday practice.

We organized the rest of the manuscript as follows. After this Introduction, we describe the dataset analyzed (Datasets), and the results we obtained (Results) Afterwards, we discuss the impact and consequences of our results, and limitations and future developments of study (Discussion), and describe the methods we employed ((Methods).

Datasets

Primary cohort and study cohort.

We analyzed a dataset made of 110,204 admissions of 84,811 hospitalized subjects between 2011 and 2012 in Norway who were diagnosed with infections, systemic inflammatory response syndrome (SIRS), sepsis by causative microbes, or septic shock[87]. The data come from the Norwegian Patient Registry[92] and the Statistics Norway agency[93].

For each patient admission, the dataset contains sex, age, septic episode number, hospitalization outcome (survival), length of stay (LOS) in the hospital, and one or more codes of the International Classification of Diseases 10th revision (ICD-10) describing the patient's disease (Table 1). Since the main goal of this study is to predict the survival of the patient, we discarded the length of stay because it strongly relates to the likelihood to survive: the longer the patient has to stay in the hospital, the less likely she/he will survive. The survival variable relates to the hospital length-of-stay, which ranges in the [0, 499] days interval and has mean of 9.351 days. Our prediction therefore refers to the likelihood of a patient to survive or decease in the 9.351 days after the collection of her/his medical record, in the hospital.

Feature	Explanation	Measurement	Range
Age	Age of the patient at the hospital stay	Years	[0, ..., 100]
Episode number	Number of septic episodes experienced by the patient	Integer	[1, ..., 5]
Sex	0: male; 1: female	Binary	0, 1
[Target] survival	0: dead; 1: alive	Boolean	0, 1

Table 1. Meanings, measurement units, and intervals of each feature of the dataset. Ranges refer both to the primary cohort and the study cohort. We used survival as prediction the target in this manuscript.

	Primary cohort		Study cohort	
Category feature	#	%	#	%
Survival (0: dead)	8105	7.35	3606	18.93
Survival (1: alive)	102,099	92.65	15,445	81.07
Sex (0: male)	57,973	52.61	10,505	55.14
Sex (1: female)	52,231	47.39	8546	44.86

Table 2. Statistical quantitative description of the category features. #: Number of admissions. %: percentage of admissions. Primary cohort full sample: 110,204 admissions. Study cohort full sample: 19,051 admissions.

	Primary cohort			Study cohort		
Numeric feature	Median	Mean	σ	Median	Mean	σ
Age	68	62.74	24.13	77	72.50	18.61
Episode number	1	1.35	0.75	1	1.40	0.75

Table 3. Statistical quantitative description of the numeric features. Primary cohort full sample: 110,204 admissions. Study cohort full sample: 19,051 admissions. σ: standard deviation.

The admissions are of 57,973 of men and of 52,231 of women, ranging from 0 to 100 years of age (Table 2 and Table 3). Most of the admissions (76.96%) relate to the first septic episode. More information about the dataset can be found in the original study[87].

The original dataset curators Knoop et al.[87] called the complete dataset with 110,204 admissions the *primary cohort*. From the primary cohort, they selected the admissions respecting the *Sepsis-3* definition of sepsis (at least one several infection or sepsis related ICD-10 code and at least one codes for acute organ dysfunction)[9], and they called this subset the *study cohort*.

Since the data of the primary cohort were recorded before the Sepsis-3 definition emerged in 2016, we cannot know if a patient diagnosed with an ICD-10 code related to sepsis actually had an organ dysfunction afterwards. Therefore, we cannot know if these admissions can be considered related to sepsis by the current Sepsis-3 definition today. What we know, instead, is that the conditions of the primary cohort (infections, systemic inflammatory response syndrome (SIRS), sepsis by causative microbes, or septic shock) might have lead to sepsis. To reflect this information, we call these conditions *sepsis potential preconditions* in this study. We decided to consider both the primary cohort and the study cohort for our analysis because consensus on a unified definition of sepsis has not been reached by the medical community yet[94].

We took the original dataset[95] and applied the same selection, generating a study cohort having a different size from the one of Knoop et al.[87]: while their study cohort contained 18,460 admissions, our study cohort contains 19,051. We were unable to obtain the original study cohort subset from Knoop unfortunately.

Both the primary cohort and our study cohort are positively imbalanced (Table 2). The primary cohort contains 102,099 admissions of patients who survived (92.65% positives) and 8,105 admissions of patients who deceased (7.35% negatives). Our study cohort contains 15,445 admissions of patients who survived (81.07% positives) and 3,606 admissions of patients who deceased (18.93% negatives).

We report the stacked barplots of sex and disease episode number (Fig. 1) and the histogram of the age distribution (Fig. 2) for the primary cohort; we report the stacked barplots of sex and sepsis episode number (Fig. 3) and the histogram of the age distribution (Fig. 4) for the study cohort.

Validation cohort.

To confirm our findings, we also applied our methods to a dataset of South Korean critically ill patients whose medical records were collected between between January 2007 and December 2015 and publically released by Lee and colleagues[96]. From their original dataset, we selected the data of 137 patients who had already 1 or 2 septic episodes.

Since all these data were recorded before 2016, they are associated to a definition of sepsis earlier than *Sepsis-3*.

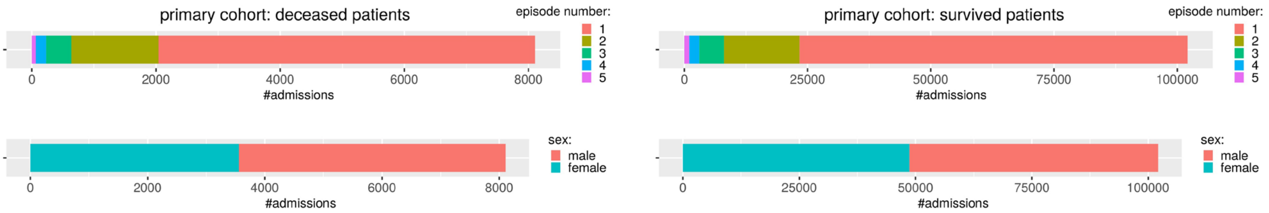

Figure 1. Primary cohort: stacked barplots of the distribution of categories. Distribution of sepsis episode number and sex of the admissions of patients who deceased (left) and survived (right). Admissions of survived patients: positives data instances (class 1). Admissions of deceased patients: negative data instances (class 0).

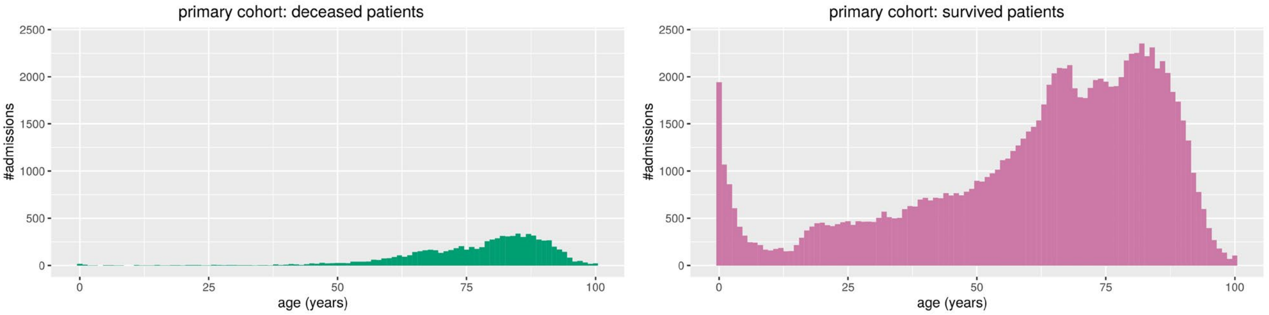

Figure 2. Primary cohort: histograms of the patients' ages in relation with the number of admissions. On the left, the admissions of the patients who deceased. On the right, the admissions of patients who survived. Admissions of survived patients: positives data instances (class 1). Admissions of deceased patients: negative data instances (class 0).

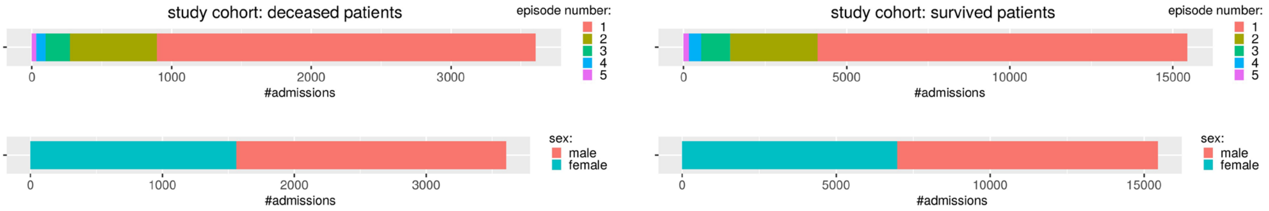

Figure 3. Study cohort: stacked barplots of the distribution of categories. Distribution of sepsis episode number and sex of the admissions of patients who deceased (left) and survived (right). Admissions of survived patients: positives data instances (class 1). Admissions of deceased patients: negative data instances (class 0).

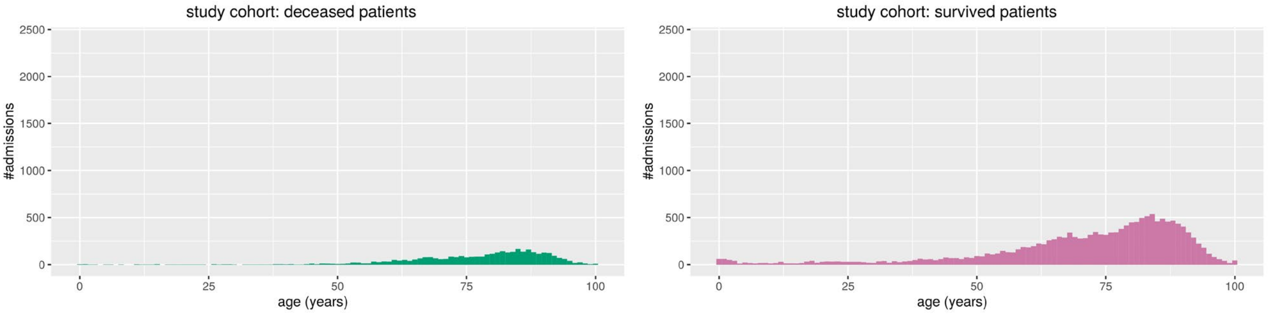

Figure 4. Study cohort: histograms of the patients' ages in relation with the number of admissions. On the left, the admissions of the patients who deceased. On the right, the admissions of patients who survived. Admissions of survived patients: positives data instances (class 1). Admissions of deceased patients: negative data instances (class 0).

The dataset already contained the sex and age features, while we deduced the septic episode feature by selecting all the patients that already had a sepsis before the surgery ("Preop shock = 1"), and devided them between the ones that had a second sepsis afterwards ("new sepsis = 1") and the ones that did not have it ("new sepsis = 0").

	Mann–Whitney	Chi-squared
Feature	Test *p* value	Test *p* value
Age	$< 2.20 \times 10^{-16}$	
Episode number	1.72×10^{-05}	
Sex		5×10^{-04}

Table 4. Results of the application of univariate biostatistics tests between each feature and the survival target, in the primary cohort. Mann–Whitney test *p*-value: probability value generated by the application of the Mann–Whitney *U* test to the corresponding feature and survival. chi-squared test *p*-value: probability value generated by the application of the chi-squared test to sex and survival. We reported the features in alphabetical order.

The 137 patients of our validation cohort are 59.54 years old on average (median: 60), 47 women and 90 men. Among them, 115 had one septic episode and 22 had two septic episodes, while 113 survived and 24 deceased. Regarding the dataset imbalance, this validation cohort is positively imbalanced, having 82.482% positive data instances and 17.518% negative data instances.

More information about this dataset can be found in the original study[96].

Results

In this section we first describe the results we obtained through the traditional univariate biostatistics tests (Statistical correlations) and then the results we achieved through the machine learning classifiers on the primary and study cohorts (Survival predictions), and on the external validation cohort (Validation on external cohort).

Statistical correlations.

We applied some traditional biostatistics tests (Biostatistics univariate tests) to evaluate univariate associations between all three feature variables and survival status on the primary cohort. Their results showed they were statistically significant with $p < 0.001$ (Table 4).

The results of these tests state there are statistically meaningful relationships between age and survival, between disease episode number and survival, and between sex and survival. These results confirm that we can use these three clinical factors as predictive features to forecast survival.

Survival predictions.

We report the results of our machine learning predictions made on the primary cohort and on the validation cohort, measured with traditional confusion matrix rates, in Table 5. As mentioned earlier, we consider positive data instances the admissions of the survived patients (class 1), and negative data instances the admissions of the deceased patients (class 0).

We report two scores considering all the possible confusion matrix thresholds (precision-recall curve and receiver operating characteristic curve), and seven scores computed by artificially setting the confusion matrix threshold to 0.5 (TP rate, TN rate, PPV, NPV, MCC, F_1 score, and accuracy). Since the main goal of our study is to predict the survived patients (positive data instances), and the inclusion of all the possible confusion matrix thresholds is more informative than the usage of an heuristic cut-off, we focused on the precision-recall area under the curve (PR AUC) as principal indicator (Table 5).

In the primary cohort, which contained admissions of patients diagnosed with sepsis before Sepsis-3, radial SVM and gradient boosting outperformed the other methods by achieving PR AUC = 0.966 and ROC AUC close to 0.7. Gradient boosting resulted being very efficient when predicting the survived patients, by achieving TP rate = 0.905, followed linear regression, that reached sensitivity = 0.805. Regarding the identification of deceased patients, the two SVM models attained the top TN rates: 0.898 for the linear SVM and 0.807 for the radial SVM.

All the five models obtained very high positive predictive values (PPVs), from linear SVM achieving 0.896 to radial SVM reaching the almost perfect value of 0.970. All the five methods, also, had low negative predictive values (NPVs), ranging from 0.112 to 0.210, which resulted in Matthews correlation coefficients, too. Regarding F_1 score and accuracy, four methods obtained high or very high results, with top performance reached by gradient boosting (F_1 score = 0.916 and accuracy = 0.851), while linear SVM achieved low scores on both these rates (Table 5).

In the study cohort, which contains admissions of patients diagnosed with sepsis based on the 2016 Sepsis-3 definition, the results were similar to those seen in the primary cohort, albeit a little lower. (Table 5). All the five models obtained very high PR AUC, with linear SVM obtaining the top score of 0.860. Regarding the ROC AUCs, the two support vector machines gained the best results with 0.568 both. Gradient boosting and linear regression were capable to correctly predict most of the survived patients, reaching sensitivity scores of 0.837 and 0.764, respectively. And linear SVM was the best at predicting deceased patients, with a specificity score of 0.898.

Regarding precision, all the five methods were capable to make accurate positive predictions, with linear SVM obtaining again the best PPV (0.896). Similar to the primary cohort, they also all had low NPV values (ranging from 0.210 to 0.239), which was reflected in their Matthews correlation coefficients. The low results on the NPVs are reflected in the Matthews correlation coefficients, too. Gradient boosting gained high values for F_1 score and accuracy also in the primary cohort (0.819 and 0.718, respectively), followed by linear regression (Table S1).

Method	PR AUC	ROC AUC	TP rate	TN rate	PPV	NPV	MCC	F_1 score	accuracy
Training and testing on the primary cohort									
Radial SVM	_0.966*_	_0.701*_	0.492	0.807	_0.970*_	0.112	_+ 0.157*_	0.652	0.515
Gradient boosting	_0.966*_	0.690	_0.905*_	0.179	0.934	0.126	+ 0.061	_0.916*_	_0.851*_
Naïve Bayes	0.954	0.649	0.553	0.745	0.965	0.117	+ 0.156	0.703	0.567
Linear regression	0.941	0.599	0.836	0.361	0.943	0.149	+ 0.135	0.886	0.801
Linear SVM	0.860	0.586	0.205	_0.898*_	0.896	_0.210*_	+ 0.104	0.333	0.337
Training and testing on the study cohort									
Linear SVM	_0.860*_	_0.586*_	0.205	_0.898*_	_0.896*_	0.210	_+ 0.104*_	0.333	0.337
Radial SVM	0.858	_0.586*_	0.408	0.718	0.861	0.222	+ 0.102	0.553	0.467
Gradient boosting	0.856	0.574	_0.837*_	0.208	0.822	0.231	+ 0.038	_0.819*_	_0.718*_
Naïve Bayes	0.841	0.562	0.405	0.718	0.861	0.220	+ 0.100	0.551	0.465
Linear regression	0.826	0.541	0.764	0.318	0.828	_0.239*_	+ 0.074	0.794	0.679

Table 5. Results of the survival prediction made with machine learning classifiers, with training phase and testing phase done on the Norwegian primary cohort or study cohort[87]. Mean results of 100 executions with random selection of the elements in the training set and test set, with ROSE oversampling[97] applied to the training set. Admissions of survived patients: positives data instances (class 1). Admissions of deceased patients: negative data instances (class 0). Linear SVM: support vector machine with linear kernel. Optimized cost regularization hyper-parameter of the linear SVM, most frequently selected C by the MCC-based grid search: $C = 0.01$ for primary cohort (63 times out of 100) and $C = 0.001$ for study cohort (51 times out of 100). Radial SVM: support vector machine with radial Gaussian kernel. Optimized cost regularization of the radial SVM, most frequently selected C by the MCC-based grid search: $C = 0.1$ for the primary cohort (56 times out of 100) and for the study cohort (51 times out of 100). MCC: Matthews correlation coefficient. MCC worst value $= -1$ and best value $= +1$. TP rate: true positive rate, sensitivity, recall. TN rate: true negative rate, specificity. PR: precision-recall curve. PPV: positive predictive value, precision. NPV: negative predictive value. ROC: receiver operating characteristic curve. AUC: area under the curve. F_1 score, accuracy, TP rate, TN rate, PPV, NPV, PR AUC, ROC AUC: worst value $= 0$ and best value $= +1$. We report the formulas of these rates in the Supplementary Information. ROSE minority class probability: $p = 0.5$ for SVMs; $p = 0.38$ for gradient boosting, naïve Bayes, and linear regression in the primary cohort; $p = 0.45$ for gradient boosting, naïve Bayes, and linear regression in the study cohort. We highlighted in italic and with an asterisk * the top result for each statistical indicator. We report the mean scores with the standard deviations in Supplementary Table S1.

Validation on external cohort. To further verify the predictive power and the generalizability of our classifiers, we performed two additional analyses involving an external validation cohort containing medical records of patients from South Korea (Datasets)[96].

In the first analysis, we both trained and tested our models on this external validation cohort, and reported the results (Table 6). In the second analysis, we trained our models on the Norwegian primary cohort or study cohort, applied the trained models to the external validation cohort, and reported the results (Table 7).

Train and test on the external validation cohort. Our results we report show that all our five methods (naïve Bayes, linear SVM, radial SVM, gradient boosting, and linear regression) are capable of efficiently predicting survival not only when trained and tested on the Norwegian cohorts, but also when trained and tested on another external dataset (Table 6). These results confirm the generalizability of our approach.

All the classifiers, in fact, obtained high PR AUC ranging from 0.873 (radial SVM) to 0.899 (linear SVM), and were able to correctly classify most of the positive data instances (minimum TP rate = 0.849) and most of the positive predictions (minimum PPV = 0.849). Only naïve Bayes and linear regression were able to correctly classify most of the negative data instances and correctly make most of negative predictions (specificity and NPV greater than 0.5 for both the methods).

Also the other indicators show good scores (ROC AUC and MCC) or optimal scores (accuracy and F_1 score for all the five classifiers, Table 6).

Train on the primary or study cohort, and test on the external validation cohort. The final part of our analysis involved the attempt to use our trained models to make survival predictions on an external dataset. In a real case scenario, in fact, physicians and medical doctors would apply our approach to the data of a new cohort of patients arriving to the hospital, and these patients of course would not be part of the original cohort where to train the models. To address this scenario, we performed an additional analysis where we trained our models on the Norwegian primary cohort or study cohort of Knoop et al.[95] and we tested them on the South Korean external validation cohort by Lee et al.[96]. We reported the results in Table 7.

As one can noticed, the algorithms we employed were able to correctly predict most of the survived patients and to make most of correct predctions, obtaining PR AUC scores ranging from 0.821 (radial SVM trained on the primary cohort) to 0.863 (gradient boosting trained on the study cohort). Naïve Bayes obtained the top score for PR AUC when trained on the primary cohort (0.848), while gradient boosting achieved the top PR AUC when trained on the study cohort (0.863). Because of the imbalance of the cohorts, all the methods achieved high

Training and testing on the validation cohort									
Method	**PR AUC**	**ROC AUC**	**TP rate**	**TN rate**	**PPV**	**NPV**	**MCC**	**F_1 score**	**accuracy**
Linear SVM	_0.899*_	0.676	0.911	0.388	0.873	0.490	+ 0.309	0.889	0.818
Naïve Bayes	0.887	_0.713*_	0.899	_0.527*_	_0.891*_	0.538	_+ 0.417*_	0.893	_0.828*_
Gradient boosting	0.883	0.682	0.912	0.448	0.885	_0.540*_	+ 0.378	_0.895*_	_0.828*_
Linear regression	0.880	0.689	0.849	0.530	0.885	0.458	+ 0.350	0.863	0.788
Radial SVM	0.873	0.642	_0.929*_	0.226	0.849	0.465	+ 0.179	0.883	0.806

Table 6. Results of the survival prediction made with machine learning classifiers on the South Korean external validation cohort[96]. Mean results of 100 executions with random selection of the elements in the training set and test set, with ROSE oversampling[97] applied to the training set. In this analysis, both the training phase and the testing phase happened on the validation cohort. σ: standard deviation. Admissions of survived patients: positives data instances (class 1). Admissions of deceased patients: negative data instances (class 0). Linear SVM: support vector machine with linear kernel. Optimized cost regularization hyper-parameter of the linear SVM, most frequently selected C by the MCC-based grid search: $C = 0.1$ (59 times out of 100). Radial SVM: support vector machine with radial Gaussian kernel. Optimized cost regularization of the radial SVM, most frequently selected C by the MCC-based grid search: $C = 0.1$ (70 times out of 100). MCC: Matthews correlation coefficient. MCC worst value = − 1 and best value = + 1. TP rate: true positive rate, sensitivity, recall. TN rate: true negative rate, specificity. PR: precision-recall curve. PPV: positive predictive value, precision. NPV: negative predictive value. ROC: receiver operating characteristic curve. AUC: area under the curve. F_1 score, accuracy, TP rate, TN rate, PPV, NPV, PR AUC, ROC AUC: worst value = 0 and best value = + 1. ROSE p-value: 0.5 for all. We report the results with standard deviations in Table S2 and the formulas of the statistical indicators in the Supplementary Information. We highlighted in italic and with an asterisk * the top result for each statistical indicator.

method	**PR AUC**	**ROC AUC**	**TP rate**	**TN rate**	**PPV**	**NPV**	**MCC**	**F_1 score**	**accuracy**
Train on primary cohort and test on validation cohort									
Naïve Bayes	_0.848*_	_0.565*_	0.715	_0.415*_	_0.852*_	_0.236*_	_+ 0.107*_	0.777	0.663
Gradient boosting	0.843	0.527	0.953	0.035	0.823	0.123	– 0.018	0.882	_0.792*_
Radial SVM	0.821	0.514	0.949	0.013	0.819	0.040	– 0.068	0.879	0.785
Train on study cohort and test on validation cohort									
Gradient boosting	_0.863*_	0.552	_0.973*_	0.061	0.830	_0.739*_	_+ 0.130*_	_0.895*_	_0.814*_
Naïve Bayes	0.848	_0.566*_	0.747	_0.386*_	_0.851*_	0.244	+ 0.113	0.795	0.683
Radial SVM	0.829	0.537	0.955	0.011	0.820	0.043	– 0.068	0.882	0.789
Linear regression	0.824	0.499	0.956	0.042	0.824	0.166	– 0.005	0.885	0.796

Table 7. Results of the survival prediction made with machine learning classifiers, including standard deviation, with training phase done on the Norwegian primary cohort or study cohort[87] and testing phase done on the South Korean external validation cohort[96]. Mean results of 100 executions with random selection of the elements in the training set and test set, with ROSE oversampling[97] applied to the training set. σ: standard deviation. Admissions of survived patients: positives data instances (class 1). Admissions of deceased patients: negative data instances (class 0). Linear SVM: support vector machine with linear kernel. Optimized cost regularization hyper-parameter of the linear SVM, most frequently selected C by the MCC-based grid search: $C = 0.01$ for primary cohort (63 times out of 100) and $C = 0.001$ for study cohort (51 times out of 100). Radial SVM: support vector machine with radial Gaussian kernel. Optimized cost regularization of the radial SVM, most frequently selected C by the MCC-based grid search: $C = 0.1$ for the primary cohort (56 times out of 100) and for the study cohort (51 times out of 100). MCC: Matthews correlation coefficient. MCC worst value = − 1 and best value = + 1. TP rate: true positive rate, sensitivity, recall. TN rate: true negative rate, specificity. PR: precision-recall curve. PPV: positive predictive value, precision. NPV: negative predictive value. ROC: receiver operating characteristic curve. AUC: area under the curve. F_1 score, accuracy, TP rate, TN rate, PPV, NPV, PR AUC, ROC AUC: worst value = 0 and best value = + 1. We report the formulas of these rates in the Supplementary Information, and the same results including the standard deviations in Table S3. ROSE minority class probability: $p = 0.5$ for SVMs; $p = 0.38$ for gradient boosting, naïve Bayes, and linear regression in the primary cohort; $p = 0.45$ for gradient boosting, naïve Bayes, and linear regression in the study cohort. We highlighted in italic and with an asterisk * the top result for each statistical indicator. We did not report the results of linear regression trained on the primary cohort and the results of the linear SVM on both the cohorts because these methods predicted all positives in the validation cohort.

scores for positive data instances (sensitivity and precision) but low scores for negative data instances (specificity and NPV). Naïve Bayes achieved the top specificity both when trained on the primary cohort (0.415) and when trained on the study cohort (0.386).

Differently from the other tests we made, some methods failed in correctly predicting any negative data instances: the linear SVM method classified all the validation set data instances as positive, both when trained on the primary cohort and on the study cohort, while the linear regression did the same for primary cohort. This aspect suggests additional future studies in the theoretical machine learning field about the behavior of these algorithms.

These results show, additionally, the level of generalizability of our approach, that is able to correctly predict survived patients just from sex, age, and septic episode even when our models are trained and tested on two different cohorts.

Discussion

Our results show that machine learning applied to minimal clinical records of patients diagnosed with sepsis, containing only age, sex, and number of septic episode, is sufficient to predict the survival outcome of the patients themselves. Most of our machine learning methods, in fact, were capable to correctly predict most of the survived patients (very high sensitivity rates) with high confidence probability (very high precision values).

To the best of our knowledge, no other study on sepsis has predicted patient survival outcomes with such little and easily obtainable information; age and sex are immediately available for each patient, while sepsis episode number can be easily found in the patient's history.

Our finding can be consequential to the way that sepsis is managed around the world. If validated, hospitals will be able to quickly and reliably predict a patient's survival in few seconds, . allowing for quicker action from the doctors, which is crucial for a quick-to-kill illness like sepsis. The finding will be especially useful to hospitals that lack personnel and machinery, like those in rural or developing areas.

Our findings were not identified by the study of the original Norwegian dataset curators, which instead provides an overall general analysis about the correlation between features of the patients' cohort[87], and not even by the study of the validation cohort[96].

As a limitation, we have to report that, even if our machine learning methods resulted being effective in identifying the admissions of survived patients, the same cannot be said for the admissions of deceased patients. Our data mining techniques, in fact, were able to correctly predict most of the admissions of the deceased patients (high TN rates), but with low diagnostic proportions (low NPVs)[98]. We believe this drawback of our study is due to the huge imbalance of the datasets: during training, the machine learning methods do not see enough negative elements, and therefore they generate many false negatives when making predictions on the test set. We tried to tackle this problem with ROSE oversampling[97], which improved the situation, but did not solve the issue. This drawback is critical because the patients who are more likely to decease are the ones who need urgent therapies and cures in a hospital setting. We hope to overcome this issue in the future by employing other oversampling techniques.

We also have to report that the absence of a temporal feature expressing the time passed between a septic episode and decease has been a limitation for this study. The presence of this time feature, in fact, would have allowed us to make time-related predictions which would have higher impact in a hospital setting, by helping doctors understanding which patients are more in need of immediate help.

In the future, we plan to further investigate the theme of the minimal clinical record for computational prediction of survival on other diseases such as cervical cancer[99], neuroblastoma[100], breast cancer[101], and amyotrophic lateral sclerosis[102].

Methods

In this section, we briefly describe the traditional biostatistics tests we employed to detect correlation between each clinical feature and survival target (Biostatistics univariate tests), and the machine learning methods we used to predict survival (Machine learning classifiers).

We implemented our software code with the free open source R programming language and platform[103], and made it publicly available online on GitHub (Data and software availability).

Biostatistics univariate tests.

To identify preliminary associations between feature (age, sex, septic episode number) and target (survival), we performed univariate biostatistics analyses. We used the Anderson–Darling test[104] to test for normality of continuous variables. As the normality assumptions were not met, we employed the Mann–Whitney U test[105] to evaluate associations between the continuous features and survival. We used the chi-squared (χ^2) test[106] to evaluate the association between sex and survival. We considered p-values less than 0.05 as statistically significant.

For both the Mann–Whitney U test and the chi-squared test, a low p-value (close to 0) means that the two analyzed features strongly relate to each other, while a high p-value (close to 1), instead, means there is no correlation[107].

Machine learning classifiers.

To predict the survival of patients from only three features, we initially employed function approximation methods[108], trying to frame this scientific problem into a linear setting, with a mathematical formula such as $y = f(x, w, z)$ where where y is survival, x is age, w is sex, and z episode number. After several attempts, however, we realized that this problem could not be solved through a simple linear function with three variables, and therefore decided to take advantage of machine learning.

We employed five machine learning classifiers from four different method families: linear regression[109], support vector machine with linear kernel (linear SVM)[110], support vector machine with radial kernel (radial SVM)[111], gradient boosting[112], and naïve Bayes[113].

We first chose linear regression because it is a baseline statistical model and one of the simplest methods in computational intelligence; starting an analysis with a simple method is considered a good practice in machine learning[114]. We then chose two support vector machines with different kernels (linear and Gaussian radial), because they can project data into a hyperplane suitable for classification. After that, we tried gradient boosting, an ensemble boosting method capable of training several weak classifiers to build a strong one. Finally, we employed a probabilistic classifier, such as naïve Bayes, which is based on the Bayesian conditional probability and can estimate how likely a data instance can belong to a class.

All these methods have shown their effectiveness in binary classification of biomedical data in the past, and therefore represented suitable candidates for this study as well.

We applied each algorithm 100 times both to the primary cohort and the study cohort and reported the mean result (Results). For methods that needed hyper-parameter optimization (linear SVM and radial SVM), we split the dataset into 60% randomly selected admissions for the training set, 20% randomly selected admissions for the validation set, and 20% remaining admissions for the test set. To choose the top hyper-parameter *C*, we used a grid search and selected the model that generated the highest Matthews correlation coefficient[114,115]. For the other methods (linear regression, naïve Bayes, and gradient boosting), instead, we severed the dataset into 80% randomly selected data instances for the training set, and 20% remaining data instances for the test set.

For each of the 100 executions, our script randomly chose admissions for the training set and for the test set (and for the validation set, in the case of hyper-parameter optimization) from the complete original primary cohort or study cohort. We trained each model on the training set (and validated it on the validation set, in the case of hyper-parameter optimization), and we then applied the model to the test set. Given the different selections of admissions for the dataset splits, each script execution generated slightly different results even when employing the same method.

Because of the huge imbalance of the datasets (92.65% positives and 7.35% negatives in the primary cohort, and 81.07% positives 18.93% negatives in the study cohort), we had to employ an oversampling technique at each execution, to make the training set more balanced. We applied the Randomly Over Sampling Examples (ROSE) method[97], which creates and adds artificial synthetic data instances of the minority class (the deceased patients, in our datasets) to the training sets. Since we split the datasets into training set, validation set, and test set for the support vector machines, and just into training set and test set for the other methods, we had to select different optimized probability values for the ROSE minority class for these two groups of algorithms.

We measured the classifiers' performances by using typical confusion matrix evaluation scores such as Matthews correlation coefficient (MCC), receiver operating characteristic area under the curve (ROC AUC), precision recall area under the curve (PR AUC), and other ones. Since our main goal is to correctly predict the survival of patients, we ranked the results based on the PR AUCs, which highlight the true positive rates and positive predictive values reached by each method[116].

Data availability

The dataset of the primary cohort and of the study cohort[95] used in this study is publicly available at: https://plos.figshare.com/articles/Epidemiology_and_impact_on_all-cause_mortality_of_sepsis_in_Norwegian_hospitals_A_national_retrospective_study/5613424. The dataset of the validation cohort[96] used in this study is publicly available at: https://figshare.com/articles/Severe_persistent_hypocholesterolemia_after_emergency_gastrointestinal_surgery_predicts_in-hospital_mortality_in_critically_ill_patients_with_diffuse_peritonitis/6770660. Our software code is publicly available at: https://github.com/davidechicco/sepsis_survival_from_age_sex_episode

Received: 19 March 2020; Accepted: 15 September 2020
Published online: 13 October 2020

References

1. World Health Organization. *WHO: Improving the Prevention, Diagnosis and Clinical Management of Sepsis*. https://www.who.int/sepsis/en/. Accessed 23 February 2020.
2. Rudd, K. E. *et al.* Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet* **395**, 200–211 (2020).
3. Nedeva, C., Menassa, J. & Puthalakath, H. Sepsis: inflammation is a necessary evil. *Front. Cell Dev. Biol.* **7**, 108 (2019).
4. Marik, P. E. The management of sepsis: science & fiction. *J. Thorac. Dis.* **12**, S1 (2020).
5. Dugar, S., Choudhary, C. & Duggal, A. Sepsis and septic shock: guideline-based management. *Clevel. Clin. J. Med.* **87**, 53–64 (2020).
6. Trovato, G. SEPSIS. Educational and best practice frontiers. Beyond the boundaries of fatality, enhancing clinical skills and precision medicine. *Ther. Clin. Risk Manag.* **16**, 87–93 (2020).
7. Leligdowicz, A. & Matthay, M. A. Heterogeneity in sepsis: new biological evidence with clinical applications. *Crit. Care* **23**, 80 (2019).
8. Arnold, C. News feature: the quest to solve sepsis. *Proc. Nat. Acad. Sci.* **115**, 3988–3991 (2018).
9. Singer, M. *et al.* The third international consensus definitions for sepsis and septic shock (Sepsis-3). *J. Am. Med. Assoc.* **315**, 801–810 (2016).
10. Horak, J., Martinkova, V., Radej, J. & Matejovic, M. Back to basics: recognition of sepsis with new definition. *J. Clin. Med.* **8**, 1838. https://doi.org/10.3390/jcm8111838 (2019).
11. Valik, J. K. *et al.* Validation of automated sepsis surveillance based on the Sepsis-3 clinical criteria against physician record review in a general hospital population: observational study using electronic health records data. *Br. Med. J. (BMJ) Qual. Saf.*, 1–18 (2020).
12. Sinha, S. & Ray, B. Sepsis-3: how useful is the new definition?. *J. Anaesthesiol. Clin. Pharmacol.* **34**, 542 (2018).
13. Rhee, C. *et al.* Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. *J. Am. Med. Assoc. Netw. Open* **2**, e187571 (2019).
14. Tao, T. *et al.* The top cited clinical research articles on sepsis: a bibliometric analysis. *Crit. Care* **16**, R110. https://doi.org/10.1186/cc11401 (2012).
15. Kempker, J. A. & Martin, G. S. A global accounting of sepsis. *Lancet* **395**, 168–170 (2020).
16. McLymont, N. & Glover, G. W. Scoring systems for the characterization of sepsis and associated outcomes. *Ann. Transl. Med.* **4**, 527–527. https://doi.org/10.21037/atm.2016.12.53 (2016).

17. Georgescu, A.-M., Szederjesi, J., Copotoiu, S.-M. & Azamfirei, L. Predicting scores correlations in patients with septic shock: a cohort study. *Rom. J. Anaesth. Intensive Care* **21**, 95–98 (2014).
18. Dabhi, A. S., Khedekar, S. S. & Mehalingam, V. A prospective study of comparison of APACHE-IV & SAPS-II scoring systems and calculation of standardised mortality rate in severe sepsis and septic shock patients. *J. Clin. Diagn. Res.* **8**, MC09-MC13 (2014).
19. Lambden, S., Laterre, P. F., Levy, M. M. & Francois, B. The SOFA score - development, utility and challenges of accurate assessment in clinical trials. *Crit. Care* **23**, 374. https://doi.org/10.1186/s13054-019-2663-7 (2019).
20. Fang, W.-F. *et al.* Application of dynamic pulse pressure and vasopressor tools for predicting outcomes in patients with sepsis in intensive care units. *J. Crit. Care* **52**, 156–162. https://doi.org/10.1016/j.jcrc.2019.05.003 (2019).
21. Li, W., Wang, M., Zhu, B., Zhu, Y. & Xi, X. Prediction of median survival time in sepsis patients by the SOFA score combined with different predictors. *Burns Trauma*https://doi.org/10.1093/burnst/tkz006 *(2020).*
22. Roh, J. *et al.* Factors predicting long-term survival of patients with sepsis on arrival at the emergency department. *Medicine* **98**, e16871. https://doi.org/10.1097/md.0000000000016871 (2019).
23. Sanderson, M. *et al.* Predicting 30-day mortality in patients with sepsis: an exploratory analysis of process of care and patient characteristics. *J. Intensive Care Soc.* **19**, 299–304 (2018).
24. Flanagan, J. R. *et al.* Predicting survival of patients with sepsis by use of regression and neural network models. *Clin. Perform. Qual. Healthc.* **4**, 96–103 (1996).
25. Mochizuki, K. *et al.* Beneficial effect modification on survival outcome of sepsis between ART-123 and polymyxin B-immobilised haemoperfusion: a nationwide Japanese registry study. *Ann. Intensive Care* **10**, 57 (2020).
26. Ferrario, M. *et al.* Mortality prediction in patients with severe septic shock: a pilot study using a target metabolomics approach. *Sci. Rep.* **6**, 20391 (2016).
27. Mewes, C. *et al.* CTLA-4 genetic variants predict survival in patients with sepsis. *J. Clin. Med.* **8**, 70. https://www.who.int/sepsis/en/3 (2019).
28. Roderburg, C. *et al.* Circulating MicroRNA-150 serum levels predict survival in patients with critical illness and sepsis. *PLoS ONE* **8**, e54612 (2013).
29. Wang, J., Sun, Y., Teng, S. & Li, K. Prediction of sepsis mortality using metabolite biomarkers in the blood: a meta-analysis of death-related pathways and prospective validation. *BMC Med.* **18**, 83 (2020).
30. Jiang, W. *et al.* Evaluation of a combination “lymphocyte apoptosis model” to predict survival of sepsis patients in an intensive care unit. *BMC Anesthesiol.* **18**, 89 (2018).
31. Luhr, R., Cao, Y., Söderquist, B. & Cajander, S. Trends in sepsis mortality over time in randomised sepsis trials: a systematic literature review and meta-analysis of mortality in the control arm, 2002–2016. *Crit. Care* **23**, 241 (2019).
32. Gwadry-Sridhar, F., Hamou, A., Lewden, B., Martin, C. & Bauer, M. Predicting sepsis: a comparison of analytical approaches. In *Proceedings of eHealth 2010—the 2010 International Conference on Electronic Healthcare, vol. 69 of Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 95–102 (Springer, Berlin, 2010).
33. Wang, S.-L., Wu, F. & Wang, B.-H. Prediction of severe sepsis using SVM model. In *Advances in Computational Biology*, 75–81 (Springer, Berlin, 2010).
34. Mani, S. *et al.* Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *J. Am. Med. Inform. Assoc.* **21**, 326–336 (2014).
35. Gultepe, E. *et al.* From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *J. Am. Med. Inform. Assoc.* **21**, 315–325 (2014).
36. Taylor, R. A. *et al.* Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad. Emerg. Med.* **23**, 269–278 (2016).
37. Desautels, T. *et al.* Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *J. Med. Intern. Res. (JMIR) Med. Inf.* **4**, e28 (2016).
38. Calvert, J. S. *et al.* A computational approach to early sepsis detection. *Comput. Biol. Med.* **74**, 69–73 (2016).
39. Horng, S. *et al.* Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE* **12**, e0174708 (2017).
40. McCoy, A. & Das, R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *Br. Med. J. (BMJ) Open Qual.* **6**, e000158 (2017).
41. Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J. & Das, R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *Br. Med. J. (BMJ) Open Respir. Res.* **4**, e000234 (2017).
42. Vellido, A., Ribas, V., Morales, C., Sanmartín, A. R. & Rodríguez, J. C. R. Machine learning in critical care: state-of-the-art and a sepsis case study. *Biomed. Eng. Online* **17**, 135 (2018).
43. Khojandi, A., Tansakul, V., Li, X., Koszalinski, R. S. & Paiva, W. Prediction of sepsis and in-hospital mortality using electronic health records. *Methods Inf. Med.* **57**, 185–193 (2018).
44. Perng, J.-W. *et al.* Mortality prediction of septic patients in the emergency department based on machine learning. *J. Clin. Med.* **8**, 1906 (2019).
45. Boyle, M. J. *Searching for Phenotypes of Sepsis: An Application of Machine Learning to Electronic Health Records*. Master’s thesis, Yale (2019). Medicine Thesis Digital Library, 3477.
46. Stanski, N. L. & Wong, H. R. Prognostic and predictive enrichment in sepsis. *Nat. Rev. Nephrol.* **16**, 20–31 (2020).
47. Islam, M. M. *et al.* Prediction of sepsis patients using machine learning approach: a meta-analysis. *Comput. Methods Progr. Biomed.* **170**, 1–9 (2019).
48. Honeyford, K. *et al.* Evaluating a digital sepsis alert in a London multisite hospital network: a natural experiment using electronic health record data. *J. Am. Med. Inform. Assoc.* **27**, 274–283 (2019).
49. Bloch, E., Rotem, T., Cohen, J., Singer, P. & Aperstein, Y. Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction. *J. Healthc. Eng.* **2019**, 1–11 (2019).
50. Choi, J.-S. *et al.* Implementation of complementary model using optimal combination of hematological parameters for sepsis screening in patients with fever. *Sci. Rep.* **10**, 1–10 (2020).
51. Ibrahim, Z. M. *et al.* On classifying sepsis heterogeneity in the ICU: insight using machine learning. *J. Am. Med. Inform. Assoc.* **27**, 437–443 (2020).
52. Delahanty, R. J., Alvarez, J., Flynn, L. M., Sherwin, R. L. & Jones, S. S. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Ann. Emerg. Med.* **73**, 334–344 (2019).
53. Wong, H. R. *et al.* Prospective clinical testing and experimental validation of the pediatric sepsis biomarker risk model. *Sci. Transl. Med.* **11**, 1–11 (2019).
54. Shashikumar, S. P. *et al.* Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *J. Electrocardiol.* **50**, 739–743 (2017).
55. Barton, C. *et al.* Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Comput. Biol. Med.* **109**, 79–84 (2019).
56. Reyna, M. A. *et al.* Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit. Care Med.* **48**, 210 (2020).

57. Le, S. *et al.* Pediatric severe sepsis prediction using machine learning. *Front. Pediatr.* **7**, 413 (2019).
58. Giannini, H. *et al.* A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit. Care Med.* **47**, 1485–1492 (2019).
59. Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* **26**, 364–373 (2020).
60. Thorsen-Meyer, H.-C. *et al.* Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit. Health* (2020).
61. Klouwenberg, P. M. C. K. *et al.* Predicting the clinical trajectory in critically ill patients with sepsis: a cohort study. *Crit. Care* **23**, 408 (2019).
62. Sheetrit, E., Nissim, N., Klimov, D. & Shahar, Y. Temporal probabilistic profiles for sepsis prediction in the ICU. In *Proceedings of KDD 2019—the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2961–2969 (Association for Computing Machinery, 2019).
63. Fagerstroem, J., Bang, M., Wilhelms, D. & Chew, M. S. LiSep LSTM: a machine learning algorithm for early detection of septic shock. *Sci. Rep.* **9**, 15132 (2019).
64. Palma, P. & Rello, J. Precision medicine for the treatment of sepsis: recent advances and future prospects. *Expert Rev. Precis. Med. Drug Dev.* **4**, 205–213 (2019).
65. Futoma, J., Hariharan, S. & Heller, K. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *Proceedings of ICML 2017—the 34th International Conference on Machine Learning*, vol. 70, 1174–1182 (2017).
66. Kamaleswaran, R. *et al.* Applying artificial intelligence to identify physiomarkers predicting severe sepsis in the PICU. *Pediatr. Crit. Care Med.* **19**, e495–e503 (2018).
67. Saqib, M., Sha, Y. & Wang, M. D. Early prediction of sepsis in EMR records using traditional ML techniques and deep learning LSTM networks. In *Proceedings of EMBC 2018—the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 4038–4041 (IEEE, 2018).
68. Lin, C. *et al.* Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. In *Proceedings of IEEE ICHI 2018—the 6th IEEE International Conference on Healthcare Informatics*, 219–228 (IEEE, 2018).
69. Brnić, M. *et al.* Sepsis prediction using artificial intelligence algorithms. In *Proceedings of IN-TECH 2018—the 9th International Conference on Innovative Technologies*, 47–50 (2018).
70. Scherpf, M., Graesser, F., Malberg, H. & Zaunseder, S. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Comput. Biol. Med.* **113**, 103395 (2019).
71. Demirer, R. M. & Demirer, O. Early prediction of sepsis from clinical data using artificial intelligence. In *Proceedings of EBBT 2019—the 2019 Scientific Meeting on Electrical-Electronics Biomedical Engineering and Computer Science*, 1–4 (IEEE, 2019).
72. Lovejoy, C. A., Buch, V. & Maruthappu, M. Artificial intelligence in the intensive care unit. *Crit. Care* **23**, 1–3 (2019).
73. Liu, R. *et al.* Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. *Sci. Rep.* **9**, 6145 (2019).
74. Nemati, S. *et al.* An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit. Care Med.* **46**, 547–553 (2018).
75. Baghaei, K. T. & Rahimi, S. Sepsis prediction: an attention-based interpretable approach. In *Proceedings of FUZZ-IEEE 2019—the IEEE International Conference on Fuzzy Systems*, 1–6 (IEEE, 2019).
76. Fleuren, L. M. *et al.* Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med.*, 1–18 (2020).
77. Levy, M. M. *et al.* Surviving sepsis campaign. *Crit. Care Med.* **43**, 3–12. (2015).
78. Rhodes, A. *et al.* The surviving sepsis campaign bundles and outcome: results from the International Multicentre Prevalence Study on Sepsis (the IMPreSS study). *Intensive Care Med.* **41**, 1620–1628 (2015).
79. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
80. Pollard, T. J. *et al.* The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
81. Knaus, W. A. & Marks, R. D. New phenotypes for sepsis: The promise and problem of applying machine learning and artificial intelligence in clinical research. *J. Am. Med. Assoc.* **321**, 1981–1982 (2019).
82. Opal, S. M. & Wittebole, X. Biomarkers of infection and sepsis. *Crit. Care Clin.* **36**, 11–22 (2020).
83. Mao, Q. *et al.* Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *Br. Med. J. (BMJ) Open* **8**, e017833 (2018).
84. Pregernig, A., Mueller, M., Held, U. & Beck-Schimmer, B. Prediction of mortality in adult patients with sepsis using six biomarkers: a systematic review and meta-analysis. *Ann. Intensive Care* **9**, 125 (2019).
85. Shukeri, W. F., Ralib, A. M., Abdulah, N. Z. & Mat-Nor, M. B. Sepsis mortality score for the prediction of mortality in septic patients. *J. Crit. Care* **43**, 163–168 (2018).
86. Dolin, H. H., Papadimos, T. J., Stepkowski, S., Chen, X. & Pan, Z. K. A novel combination of biomarkers to herald the onset of sepsis prior to the manifestation of symptoms. *Shock* **49**, 364–370 (2018).
87. Knoop, S. T., Skrede, S., Langeland, N. & Flaatten, H. K. Epidemiology and impact on all-cause mortality of sepsis in Norwegian hospitals: A national retrospective study. *PLoS ONE* **12**, e0187990 (2017).
88. Ginestra, J. *et al.* Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Crit. Care Med.* **47**, 1477–1484 (2019).
89. Joshi, M. *et al.* Digital alerting and outcomes in patients with sepsis: systematic review and meta-analysis. *J. Med. Intern. Res.* **21**, e15166 (2019).
90. United States Food and Drugs Administration (FDA). *Software as a Medical Device (SaMD)*. https://www.fda.gov/medical-devices/digital-health/software-medical-device-samd. Accessed 23 February 2020.
91. Cheney, C. *HealthLeadersMedia.com—Spot on: New Decision Support Tool Reduces Sepsis Mortality by 22.9%*. https://www.healthleadersmedia.com/clinical-care/spot-new-decision-support-tool-reduces-sepsis-mortality-229. Accessed 23 February 2020.
92. Norwegian Patient Registry. *Norsk Pasientregister (NPR)*. https://helsedirektoratet.no/english/norwegian-patient-registry. Accessed 24 February 2020.
93. Statistics Norway. *Statistisk sentralbyra*. https://www.ssb.no/en. Accessed 24 February 2020.
94. Saria, S. & Henry, K. E. Too many definitions of sepsis: can machine learning leverage the electronic health record to increase accuracy and bring consensus?. *Crit. Care Med.* **48**, 137–141 (2020).
95. Knoop, S. T., Skrede, S., Langeland, N. & Flaatten, H. K. *Epidemiology and Impact on All-cause Mortality of Sepsis in Norwegian Hospitals: A National Retrospective Study—S1 Dataset*. https://plos.figshare.com/articles/Epidemiology_and_impact_on_all-cause_mortality_of_sepsis_in_Norwegian_hospitals_A_national_retrospective_study/5613424 (2017).
96. Lee, S. H. *et al.* Severe persistent hypocholesterolemia after emergency gastrointestinal surgery predicts in-hospital mortality in critically ill patients with diffuse peritonitis. *PLoS ONE* **13**, e0200187 (2018).
97. Lunardon, N., Menardi, G. & Torelli, N. ROSE: a package for binary imbalanced learning. *R J.* **6**, 79–89 (2014).
98. LaMorte, W. W. *Screening for Disease: Positive and Negative Predictive Value*. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_Screening/EP713_Screening5.html. Accessed 24 February 2020 (2016).

99. Fernandes, K., Chicco, D., Cardoso, J. S. & Fernandes, J. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. *PeerJ Comput. Sci.* **4**, e154 (2018).
100. Maggio, V., Chierici, M., Jurman, G. & Furlanello, C. Distillation of the clinical algorithm improves prognosis by multi-task deep learning in high-risk neuroblastoma. *PLoS ONE* **13**, e0208924 (2018).
101. Patrício, M. *et al.* Using resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer* **18**, 29 (2018).
102. Kueffner, R. *et al.* Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Sci. Rep.* **9**, 1–14 (2019).
103. The R Project for Statistical Computing. What is R? https://www.r-project.org/. Accessed 24 February 2020.
104. Anderson, T. W. & Darling, D. A. A test of goodness of fit. *J. Am. Stat. Assoc.* **49**, 765–769 (1954).
105. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
106. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *J. Sci.* **50**, 157–175 (1900).
107. Chicco, D. & Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med. Inform. Decis. Mak.* **20**, 1–16 (2020).
108. Esch, R. Functional approximation. In *Handbook of Applied Mathematics*, 928–987 (Springer, 1990).
109. Seber, G. A. & Lee, A. J. *Linear Regression Analysis, Vol. 329 of Wiley Series in Probability and Statistics* (Wiley, Hoboken, 2012).
110. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
111. Amari, S.-I. & Wu, S. Improving support vector machine classifiers by modifying kernel functions. *Neural Netw.* **12**, 783–789 (1999).
112. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
113. Rish, I. An empirical study of the naive Bayes classifier. In *Proceedings of IJCAI 2001—the 17th International Joint Conferences on Artificial Intelligence Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, 41–46 (2001).
114. Chicco, D. T. quick tips for machine learning in computational biology. *BioData Min.* **10**, 1–17 (2017).
115. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**, 1–13 (2020).
116. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).

Acknowledgements

The authors thank Christian Cumbaa and Max Kotlyar (Krembil Research Institute) for their helpful suggestions, and Zuhaib Ahmed (Krembil Research Institute) for his English proof-reading.

Author contributions

D.C. conceived the study, found the datasets, implemented the methods, and conducted the tests. D.C. and G.J. analyzed the results. G.J. did the literature search, wrote the introduction, and supported the cost of the publication. Both authors reviewed the manuscript and approved its final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-73558-3.

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

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.