Computer Methods and Programs in Biomedicine 188 (2020) 105295

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Kernel density estimates for sepsis classification

Jacquelyn Dawn Parente[a], J. Geoffrey Chase[b,*], Knut Möller[a], Geoffrey M. Shaw[c]

[a] Institute of Technical Medicine, Furtwangen University, Villingen-Schwenningen, Germany
[b] Centre of Bioengineering, University of Canterbury, Christchurch, New Zealand
[c] Intensive Care Unit, Canterbury District Health Board, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 18 September 2019
Revised 19 December 2019
Accepted 21 December 2019

Keywords:
Kernel density
Classification
Sepsis
Intensive care

ABSTRACT

Objective: Severe sepsis is a leading cause of intensive care unit (ICU) admission, length of stay, mortality, and cost. systemic inflammatory response syndrome (SIRS) and organ failure due to infection define it, but also make it hard to diagnose. Early diagnosis reduces morbidity, mortality and cost, and diagnosis is often significantly delayed due to a lack of effective biomarkers. This research employs kernel density estimation (KDE) methods fusing a personalized, model-based insulin sensitivity (SI) metric with standard bedside measures of: temperature, heart rate, respiratory rate, systolic and diastolic blood pressure, and SIRS, as these measures are available hourly or more frequently.

Methods: Model-based SI is a derived metric, identified using clinical data and a clinically validated metabolic model. The KDE classifier discriminates severe sepsis and septic shock from moderate sepsis using accepted consensus sepsis scores. A best case in-sample estimate, a worst case independent cross validation estimate, and an accepted .632 bootstrap estimate are calculated to assess performance using multi-level likelihood ratios, and sensitivity and specificity. Performance is assessed against clinically and statistically defined thresholds denoted for the minimum acceptable level as: "*high accuracy, often providing useful information, and clinical significance*," and similar definitions for greater or lesser quality.

Results: The .632 bootstrap estimate performs near clinically defined levels of high accuracy, often providing useful information, and clinical significance based on sensitivity, specificity, and multilevel likelihood ratios.

Conclusion and significance: The classifier created and this overall approach is useful for clinical decision making in diagnosing severe sepsis and septic shock in real time, for both case and control hours. However, improvements could be made with larger clinical data sets.

1. Introduction

Severe sepsis is defined by infection associated with systemic inflammatory response syndrome (SIRS) and organ failure [3,24]. Severe sepsis has 11–15% incidence, 30–60% mortality, US$16.7B annual total cost, 1.5% projected annual increase [1], and is the 11th leading cause of death in the US [32]. Early diagnosis and treatment are critical [22]. However, blood cultures to identify the presence of infection take 1–2 days so treatment has to begin before diagnosis to reduce negative outcomes.

Due to its definition, sepsis diagnosis is thus difficult, if not impossible, in real-time. As a result, 50% of sepsis cases are deemed "culture negative" [15,28,35], and diagnosis and early care to guidelines are based primarily on clinical experience and intuition balanced against the risk of not treating sepsis early enough. Care thus becomes more variable and the potential overuse of antibiotics arises in treatment without diagnosis or "culture negative" cases to mitigate risk. There is thus a need for objective, accurate, real-time diagnostic tools for sepsis.

Some sepsis diagnosis biomarkers have been assessed, but none have provided high enough sensitivity or specificity for regular use in care [36]. In particular, hourly detection methods would provide a more rapid approach, and enable appropriate antibiotic dosing at the earliest opportunity, which has been shown to reduce mortality. Currently, no such diagnostic method exists.

Kernel density estimation (KDE) offer another approach to classify, identify and diagnose potential biomarkers for diagnosis [18,31]. KDEs provide smooth, continuous probability density functions predicting the distribution of a random variable drawn from

* Corresponding author.
E-mail addresses: pjd@hs-furtwangen.de (J.D. Parente), geoff.chase@canterbury.ac.nz (J.G. Chase), moe@hs-furtwangen.de (K. Möller), geoff.shaw@health.govt.nz (G.M. Shaw).

https://doi.org/10.1016/j.cmpb.2019.105295
0169-2607/

a finite sample. They are non-parametric and using Bayes' theorem provide a posterior probability to enable classification [18]. Thus, KDE provides a probabilistic diagnostic approach given an objective metric associated with the disease state.

Model-based insulin sensitivity (S_I) is one potential biomarker or metric. It has demonstrated a strong relation to patient condition, is personalized to patient state in real-time, and is objectively calculated [17]. This research combines it with standard, readily available clinical metrics used to diagnose sepsis. A kernel implementation of the Bayes classifier is presented, which estimates class conditional densities of severe sepsis and septic shock cases versus SIRS and sepsis controls, enabling classification and thus diagnosis.

This insulin sensitivity (S_I) is identified from clinical data in real time [5,8], and decreases with worsening condition [2,4,25,37,43], while increasing as patient condition improves [2,12,23,25,37,43]. S_I alone identifies 75% of patient hours with sepsis [2], as high S_I accurately rules out sepsis. Adding further available physiological data in diagnosing severe sepsis yields 73% sensitivity and 80% specificity [25]. However, ruling out sepsis using this approach is not the same as diagnosing sepsis.

Mica et al. used density estimates to assess the diagnostic quality of trauma scores for SIRS and sepsis in polytrauma patients [30]. However, density estimates were not used for classification or diagnosis, but to supplement statistics on incidence of each level of sepsis. They showed APACHE II scores [21] distinguished no SIRS and sepsis with moderate accuracy (0.82 (0.73–0.88) AUC), while all other trauma scores had low accuracy. However, the clinically relevant question—discrimination between SIRS and sepsis—was not reported.

Martínez-Camblor explored the impact of classification errors on clinical decision making and the effects of variability [27]. Kernel density estimates were used to assess procalcitonin (PCT) levels as a diagnostic biomarker, resulting in 0.88 sensitivity and 0.80 specificity. However, the authors did not provide the sepsis definition used, population data, PCT cutoff levels, or sample selection criteria, making repeatability, analysis, and comparison to other studies difficult [14,33,42,44], further justifying the work in this research.

2. Methods

2.1. Clinical data

This case-control study compared the physiological symptoms of cases (severe sepsis and septic shock) and controls (SIRS and sepsis) in real-time. 10,048 h of sample data were obtained from the patient records of 36 adults in the Christchurch Hospital ICU with confirmed sepsis in while on the SPRINT glycemic control protocol [6]. Hourly clinical data includes: model-based insulin sensitivity (S_I) identified using clinical data and the ICING model [26], heart rate, temperature, systolic and diastolic blood pressure, respiratory rate, and SIRS score. S_I is clincally validated and employed in glycemic control [12,39,40]. Approval of this study and the use of this data was granted by the NZ Upper South Islands Ethics Committee.

ACCP/SCCM sepsis definitions [24] were applied to patient data to categorize each hour (1–4): as SIRS, sepsis, severe sepsis, or septic shock. Patient hours were removed if they had missing data, identified $S_I = 0$, or were without infection and less than two SIRS criteria. Finally, there were a total of 6071 patient hours available.

Cohorts were defined at a discrimination level of severe sepsis. In this data set, samples comprised 213 h of severe sepsis and septic shock cases and 5858 h of SIRS and sepsis controls Table 1. Sepsis prevalence was 3.5%, which is within clinical reported ranges. Fig. 1 plots the physiological data by sepsis score, and indicates some variables may be more discriminatory across sepsis levels. The KDE classifier was thus designed to discriminate severe sepsis ((3) and septic shock (4), the clinically higher impact levels, from SIRS (1) and sepsis (2) controls using this readily available data. It thus descriminates between significant sepsis warranting specialized treatment [24] and severe illness.

Table 1
Hours (%) of patient state: SIRS, sepsis, severe sepsis, and septic shock [24].

	SIRS	Sepsis	Severe sepsis	Septic shock
Raw	4918 (48.95)	4888 (48.65)	91 (0.91)	151 (1.50)
Filtered	1558 (25.66)	4300 (70.83)	85 (1.40)	128 (2.11)

2.2. Performance assessment

Diagnostic test accuracy is assessed using several recommended measures [13]:

- Likelihood ratios (LHR).
- Multilevel likelihood ratios (MLR).
- Receiver operating characteristic (ROC) curve.
- Area under the ROC curve (AUC).
- ROC cutoff yielding the highest discriminative ability.
- Confidence intervals for each measure.

Performance and accuracy are assessed by the *potential to alter clinical decisions*. For likelihood ratios (LHR+ and LHR-), these tests have positive and negative likelihood ratios LHR+ $>$ 10 and LHR– $<$ 0.1. Tests which *often provide useful information*, but do not alter clinical decisions, are defined by LHR+ of 5–10, and LHR– of 0.1–0.2. Tests with LHR+ $<$ 3 and LHR– $>$ 0.33 *rarely alter clinical decisions*. All of these definitions are given in detail in [19]. Similarly, *potentially useful tests* have a diagnostic odds ratio (DOR) well above 20 [13]. Finally, with respect to AUC, *perfect tests* have AUC = 1.0. AUC $>$ 0.9 has *high accuracy*, AUC = 0.7–0.9 is *moderate accuracy*, and AUC = 0.5–0.7 is *low accuracy* [41].

These test measures thus account for a wide range of assessment and the wide range of incidence in Table 1.

Additional measures include: sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Sensitivity and specificity over 0.9 or 90% are sufficient to be routinely employed in clinical practice [36]. PPV and NPV are strongly influenced by infection prevalence [38], where the very low incidence of severe sepsis makes high PPV unlikely. Hence, this research employs tests independent of prevalence (LHR, AUC, and DOR) [13], where the wide range used provides robust evaluation over several approaches.

2.3. Overall approach

KDEs were used to develop joint probability density profiles for 213 h of severe sepsis and septic shock cases and 5858 h of SIRS and sepsis controls, and for classification. A kernel probability density profile was made for each cohort and for the clinical predictor. Thus, a single density is used to encompass the predictors. Finally, the unknown patient hour to be classified was tested against these established datasets, with the result being a classification into either the case or control group. Optimal diagnostic performance from the ROC curve was determined for resubstitution [18], bootstrap [11], and .632 bootstrap estimates [10], which provide best case, worst case, and likely case outcome estimates.

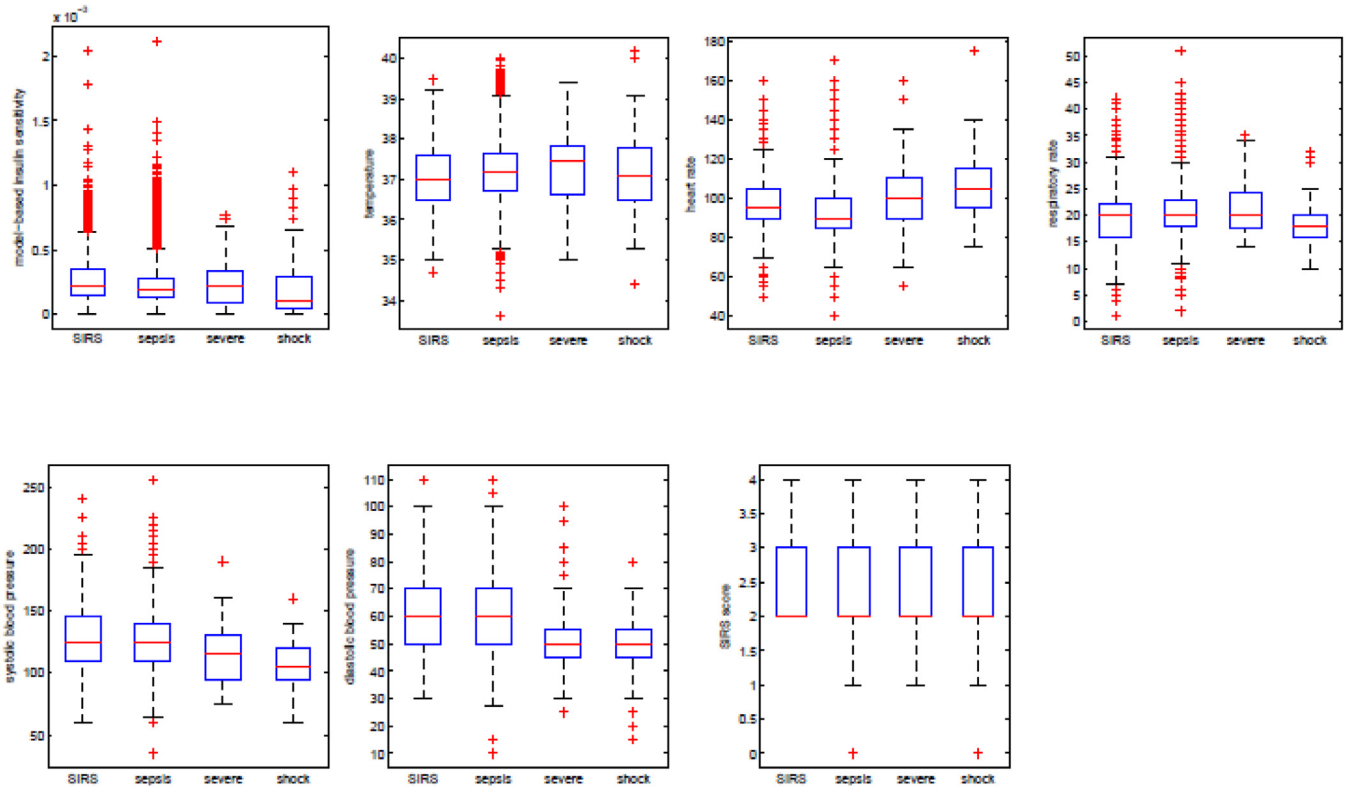

Fig. 1. Box and whisker plots of clinical data by sepsis level. Clincal data comprises: model-based insulin sensitivity, temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, and SIRS score. The four sepsis levels are 1–(4): SIRS, sepsis, severe sepsis, and septic shock.

2.4. Kernel density estimates

2.4.1. Classification

Classification problems are defined as determining in which class a set of observed data belong. A classifier is a decision rule assigning the data to a class identity. Using a Bayes classifier, each observation is assigned to the class with the largest posterior probability. Thus, for this binary sepsis classification case, there are two vectors (x) of observed data hours (M) of predictor dimensions (d) from the cases (S: severe sepsis and septic shock) and control (N: SIRS and sepsis) classes.

Let x_0^* denote the values of the clinical predictors at the given patient hour, and $\hat{f}_S(x_0^*)$ and $\hat{f}_N(x_0^*)$ denote the joint probability densities for cases and controls at that value. For each hour, the posterior probability of belonging to cases (S), given the data values obtained at that hour is defined [18]:

$$\hat{Pr}(S|x_0^*) = \frac{\hat{\pi}_S \hat{f}_S(x_0^*)}{\hat{\pi}_S \hat{f}_S(x_0^*) + \hat{\pi}_N \hat{f}_N(x_0^*)} \tag{1}$$

The posterior probability of belonging to the controls (N), given data values obtained that hour is $\hat{Pr}(N|x_0^*) = 1 - \hat{Pr}(S|x_0^*)$.

In Eq. (1), $\hat{\pi}_S$ and $\hat{\pi}_N$ are the prior probabilities of the sample being in that group. The prior probabilities are usually set to the sample proportions [18] or to 0.5 at that hour, if there is no reliable information or clinical judgement is a chance result. Thus, if the ratio in Eq. (1) is greater than a specified threshold, the sample is classified as group S, otherwise it is group N.

2.4.2. Kernel density estimation

Kernel density estimation is a non-parametric estimation of the probability density function of a random variable. Study data are treated as samples of independently and identically distributed random variables, from a distribution with unknown density. The goal is to estimate the shape of this density function. A kernel implementation of a Bayes classifier estimates class conditional densities using kernel density estimators [18].

As the class identities are known for each hour of this data, they can be used as a training sample to construct a classifier. Kernel density estimation is used to estimate f_S and f_N, the conditional probability functions of $x \in S$ given data on S and $x \in N$ given data on N. The kernel density estimator for $\hat{f}_S$ is defined:

$$\hat{f}_S(x) = \hat{f}(x|S) = \frac{1}{M_S} \sum_{i=1}^{M_S} \phi(x; x_{S,i}, H_S) \tag{2}$$

where M_S is the total number of case hours, φ is a d-variate normal density with mean vector $x_{S,i}$ and $d \times d$ covariance matrix and bandwidth matrix $H_S \in \mathbb{R}^{d \times d}$. For example, if $d = 2$, it results in a contour plot of bivariate normal density with non-zero diagonal covariance elements. In this study, d=7, which is the total number of clinical predictors employed.

2.4.3. Practical considerations

The natural logarithm is used to transform positive data values. The vector x is orthogonalized using the Cholesky or PCA transform to rotate and scale the observed predictor vectors, making the covariance matrix for the orthonormalized x equal to the identity matrix. This orthonormalisation provides matrix stability and a fixed reference value.

2.4.4. Product kernel

Assuming a diagonal bandwidth matrix ($H_i = diag(h_{i,1}, ..., h_{i,d})$), which presumes variable independence, φ becomes:

$$\phi(x; x_{i,j}, H_i) = \prod_{\ell=1}^{d} \phi(x_\ell; x_{i,j,\ell}, h_{i,\ell}) \tag{3}$$

where φ is univariate $N(x_{i,j,\ell}, h^2_{i,\ell})$, which reduces the number of bandwidth components needed to specify to d components. Thus, the kernel estimator for the joint density or class conditional probability can be written:

$$\hat{f}_S(x) = \frac{1}{M_S}\sum_{i=1}^{M_S}\left(\prod_{\ell=1}^{d}\phi\left(x_\ell; x_{S,j,\ell}, h_{S,i}\right)\right) \tag{4}$$

where the bandwidth $h_{S,i} = min\{s_{S,i}, \frac{IQR_{S,i}}{1.348}\}\ M_S^{-1/(4+d)}$, $s_{S,i}$ is the standard deviation and *IQR* the interquartile range for the component. This approach results in a product of univariate kernels, and thus the name, product kernel (PK) [7].

The kernel product assumes the components of x are independent. PK does not assume this independence, but, instead, makes the weaker assumption the kernel has independent components. Practically, these assumptions mean the various clinical measurements used as predictors are assumed independent, even if they are all associated with sepsis.

2.4.5. Implementation and comparison

Overall, the calculations allow implementation of prior probability (sample population or 0.05), vector transformation (Cholesky or PCA), and PK or kernel product. All available permutations were implemented to calculate the joint probabilitiy densities for application into Eq. (1). Then the best estimate is selected based on performace assessment. We assess performance as discussed in Section 2.5.

More specifically, the creation and implementation of the classifier takes the following steps:

- *Kernel Density Construction:* The kernel density is constructed per Eqs. (3) and (4) using the clinical data and sepsis scores.
- *Classification:* Any given hour of clinical data can be used to estimate the probability of sepsis using the classifier in Eq. (1).

This two-step process relies on the input data used and the kernel density of Eq. (4) will vary with the input clinical data used, thus changing the classification probability in Eq. (1) for any given hour.

This latter point is critical in validation, where cross validation and performance assessment via bootstrap estimation uses 80% of the data to build these kernel density in Eq. (4), and then tests on the remaining 20% of the data, where the data proportions are randomly selected and the process is repeated 1000 times to ensure robustness and that no subset of data is dominant. This bootstrap estimate provides a worst case estimate and cross validation of the kernel density estimate method. Equally, a resubstitution estimate, the best case, uses all the data to create the kernel density of Eq. (4) and tests on the same data, providing a best case result

2.5. Validation and kernel density estimator performance assessment

The validation of each estimate and its performance are assessed in the following 3 ways.

2.5.1. Resubstitution estimate

The resubstitution estimate is obtained using the same sample to construct the classifier and to assess its performance. Hence, it underestimates the true error rate because it has been developed or trained and tested using the same data. Resubstitution estimates thus represent the best case or maximum classifier performance and provide a validation based on no independent test data.

2.5.2. Bootstrap estimate

Cross-validation is a typical method to offset the best case estimate of the resubstitution estimate, but can yield high variability despite small bias. To reduce the high variability of cross-validation, bootstrap estimators were proposed [11]. Bootstrap estimates using the stratified bootstrap method randomly remove 20% of the data to test the classifier. In this study, 1000 bootstrap runs were used to estimate mean classification error. Bootstrap estimates thus represent a measure of the minimum performance of the developed classifier, and is a form of 5-fold cross validation with Monte Carlo analysis repeating it 1000 times to test on 20% of the data being independent.

2.5.3. The .632 bootstrap estimate

Since each bootstrap sample of size n has only .632n different observations on average [11], the bootstrap estimate tends to overestimate the true error rate. Thus, [11] proposed the weight of .632 to mitigate this overestimation. Thus, the .632 bootstrap estimate represents the overall performance of the developed classifier, a bias corrected estimate between best (resubstitution) and worst (bootstrap) estimator cases.

2.5.4. Overall validation and performance assessment summary

All 3 estimates provide maximum (best case; resubstitution estimate), minimum (worst case; bootstrap estimate), and overall performance (expected case; .632 bootstrap estimate) assessments. There are thus 3 performance measures and a best case and worst case validation of the accuracy of these estimators. From these three possible values we compare AUC and likelihood values to find the best case KDE.

These metrics assess the performance of the kernel density method proposed, but also assess and validate the performance relative to the number of patients. Specifically, the bootstrap estimate is a Monte Carlo approach to 5-fold cross validation, using 80% of the data (randomly selected without replacement) to build the kernel and the remaining 20% of the data to test the kernel. It is repeated 1000 times ensuring a complete and robust analysis and validation.

As noted, the resubstitution estimate, where all data is used to create and test the kernel density estimator, underestimates the data and is a “best case”. The bootstrap estimation is a cross validation method, but tends to overestimate error and is a worst case estimate. Thus, the 0.632 bootstrap estimate mixes these results for a best case error. Hence, the estimation performance assessment “worst case” error from the bootstrap estimate is a rigorous cross validation test. More specifically, cross validation is built into this performance assessment error metric and tests the classifier on “hidden data” in each Monte Carlo run and estimate.

3. Results

3.1. Kernel density estimates

PK estimates using Eq. (3) produced the greatest resubstitution AUC = 0.98–0.99 values, outperforming all kernel product estimates with AUC = 0.81–0.85. Disease prevalence strongly skewed the distribution of posterior probabilities, while 0.5 priors were not skewed and provided higher AUC values for all error estimates. PCA transformations were unstable for bootstrap estimates, but Cholesky transformations performed well. Thus, the following results use the most stable and accurate estimtes. Specifically, PK, with 0.5 for prior probabilities and Cholesky transformation.

3.2. Resubstitution estimate

At an optimal probability cutoff value of 0.35, the resubstitution estimate had 94% sensitivity, 94% specificity, 0.99 AUC, 15.70 LHR+, 0.06 LHR–, 0.36 PPV, 1.00 NPV, and 262 DOR (Table 2). This level of AUC performance is *highly accurate* [41], while the DOR shows this test is *potentially useful* [13]. Sensitivity and specificity perform at clinically significant levels *sufficient to be routinely employed in clinical practice* [36]. Positive test results are obtained 15.7 times more

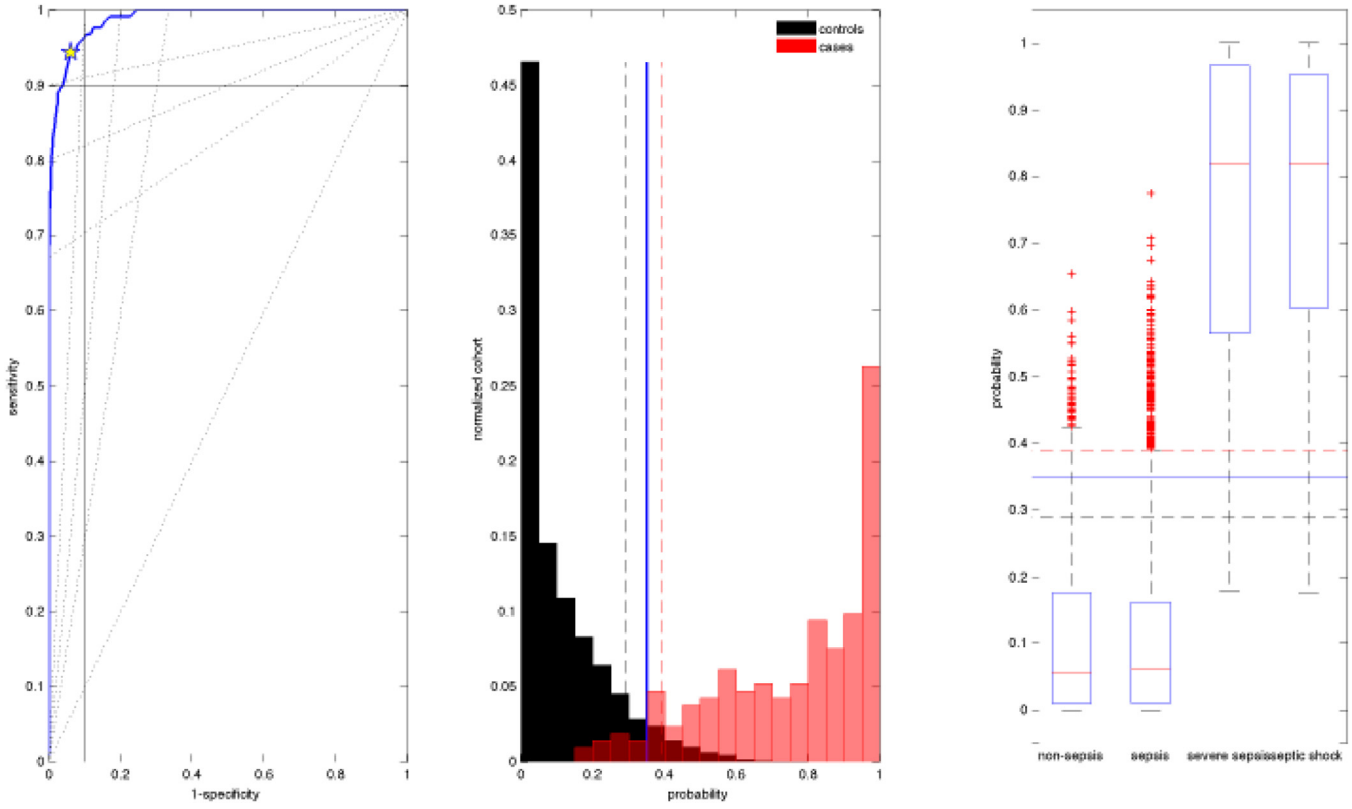

Fig. 2. Subplot 1: ROC curve for the resubstitution estimate, the best case error. Subplot 2: Histogram of posterior probabilities normalized by cohorts. Subplot 3: Box and whisker plot of posterior probabilities by sepsis score.

Table 2
Resubstitution estimate contingency table results.

Cutoff	Cases	Controls	Predictive
0.35	213	5858	**Values**
Positive	201	352	0.36 PPV
Negative	12	5506	1.0 NPV
Performance	0.94 sensitivity	0.94 specificity	0.99 AUC
Likelihood	15.70 LHR+	0.06 LHR–	262 DOR

Table 3
Resubstitution estimate LHR regions and MLRs. LHR regions with the potential to alter clinical decisions occur at cutoff values of 0.3 or greater for LHR+ and 0.38 and less for LHR– (values of LHR+ > 0.2 and LHR– < 0.54 often provide useful information. Finally, LHR+ < 0.13 and LHR– > 0.66 rarely alter clinical decisions [19]).

LHR–	≤ 0.1	0.2	0.33
Cutoff	0–0.38	0.54	0.66
Probability	**Cases**	**Controls**	**LHR–**
0.00–0.08	0	3282	0.00
0.08–0.17	0	1152	0.00
0.17–0.26	5	700	0.20
0.26–0.35	7	372	0.52
LHR+	3	5	≥ 10
Cutoff	0.13	0.2	0.3–1
Probability	**Cases**	**Controls**	**LHR+**
0.35–0.51	24	286	2.31
0.51–0.67	37	62	16.41
0.67–0.83	37	4	254.40
0.83–1.00	103	0	Inf

often from a case hour than a control hour (LHR+), while negative test results are less than six-one-hundredths as likely to be found in a case hour than from a control hour (LHR–). Thus, both LHRs have the *potential to alter clinical decisions* [19]. However, resubstitution is a best case estimate.

Fig. 2 shows the ROC curve, probabilities normalized by cohort, and probabilities by sepsis score. For the ROC curve, clinically significant 90% sensitivity is reached at cutoff value 0.39 with 96% specificity. Similarly, 90% specificity is obtained at 0.29 with 96% sensitivity. The histogram shows the optimal cutoff value and 90% sensitivity and specificity cutoff values, with strong discrimination cases (S) and controls (N). The box and whisker plot shows increasing sepsis severity does not alter specificity (0.93 and 0.94) or sensitivity (0.93 and 0.95), yielding near perfect accuracy independent of severity of sepsis.

Table 3 shows probability values obtained above the optimal cutoff value *often provide useful additional information* and have the *potential to alter clinical decisions* [19]. The greater the probability, the greater the LHR+. Thus, values above 0.51 are very likely from case hours. Alternatively, LHR– < 0.26 are very likely from control hours. Thus, MLR values in Table 3 show useful information for the positive identification of cases with greater accuracy with increasing probabilities and for the correct identification of controls with greater accuracy with decreasing probabilities.

3.3. Bootstrap estimate

At optimal probability of 0.30, Table 4 shows the bootstrap estimate yielded: 69% sensitivity, 76% specificity, AUC = 0.78, 2.88 LHR+, 0.41 LHR–, 10% PPV, 99% NPV, and DOR = 7.04. The classifier identifies the majority of both control and case hours, but is not clinically significant [36]. LHRs perform at levels that *rarely alter clinical decisions* [19]. Although the bootstrap estimate represents

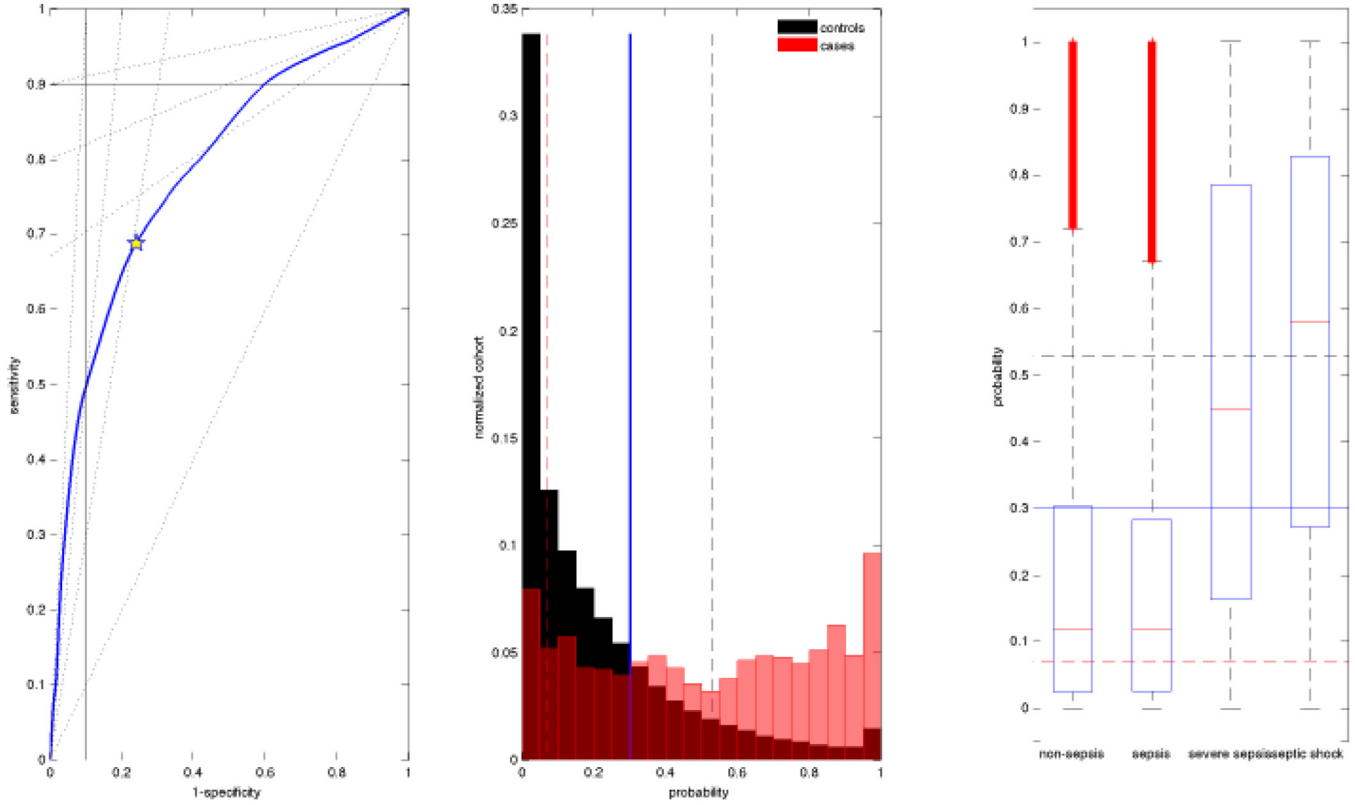

Fig. 3. Subplot 1: ROC curve for bootstrap estimates. Subplot 2: Histogram of probability normalized by cohorts. Subplot 3: Box and whisker plot of probability by sepsis score.

Table 4
Contingency table for bootstrap estimates

Cutoff	Cases	Controls	Predictive
0.30	43,000	1,172,000	**Values**
Positive	29,585	279,637	0.10 PPV
Negative	13,415	892,363	0.99 NPV
Performance	0.69 Sensitivity	0.76 Specificity	0.78 AUC
Likelihood	2.88 LHR+	0.41 LHR–	7.04 DOR

minimum classifier performance, the AUC shows *moderate accuracy* [41].

Fig. 3 shows the results. In the ROC curve, 90% sensitivity is reached at cutoff 0.07 with 39% specificity, and 90% specificity is at 0.53 with 50% sensitivity. The histogram shows overlap between cases and controls. The box and whisker plot shows increasing sepsis severity does not change specificity (0.75 and 0.77), but does change sensitivity (0.63 and 0.73). Thus, the bootstrap estimate yields a moderate test result with overlap between cases and controls. However, it is the minimum performance or worst case.

Table 5 shows LHR+ regions with the potential to alter clinical decisions occur at a cutoff value of 1. LHR– regions do not perform in ranges contributing to clinical decision making, where cutoff values greater than 0.14 for negative results rarely alter clinical decisions [19]. MLRs in Table 5 show positive results at probability values above 0.64 often provide useful additional information [19] to identify cases. However, probability values obtained below the optimal cutoff value rarely alter clinical decisions [19]. MLRs for negative results at probability values below the optimal cutoff value perform at values that rarely alter alter clinical decisions [19]. Thus, MLR values for bootstrap estimates show even the worst-case estimate often provides useful information for the positive identification of cases.

Table 5
Bootstrap esimate LHR regions and MLRs.

LHR–	$\leq$ 0.1	0.2	0.33
Cutoff	–	–	0.14–1
Probability	Cases	Controls	LHR–
0.00–0.07	4134	460,723	0.24
0.07–0.15	3961	197,017	0.55
0.15–0.22	2571	126,781	0.55
0.22–0.30	2749	107,842	0.69
LHR+	3	5	$\geq$ 10
Cutoff	0.32	0.53	1
Probability	Cases	Controls	LHR+
0.30–0.47	6536	134,479	1.32
0.47–0.64	5444	69,147	2.15
0.64–0.82	7301	39,915	4.99
0.82–1.00	10,304	36,096	7.78

Table 6
Contingency table for .632 bootstrap estimates.

0.31 Cutoff	Cases	Controls	
Performance	0.78 sensitivity	0.83 specificity	0.87 AUC
Likelihood	4.48 LHR+	0.27 LHR–	16.83 DOR

3.4. The .632 bootstrap estimate

Table 6 shows results for the .632 bootstrap estimate with an optimal cutoff value of 0.31 yielding 78% sensitivity, 83% specificity, AUC = 0.87, 4.48 LHR+, 0.27 LHR–, and DOR = 16.83. It identifies the majority of both case and control hours, but sensitivity and specificity do not reach the 90% clinically significant threshold [36]. LHR performance is indeterminate, where both LHRs are outside of the range of rarely altering clinical decisions, yet are not yet within the range of often providing useful information [19].

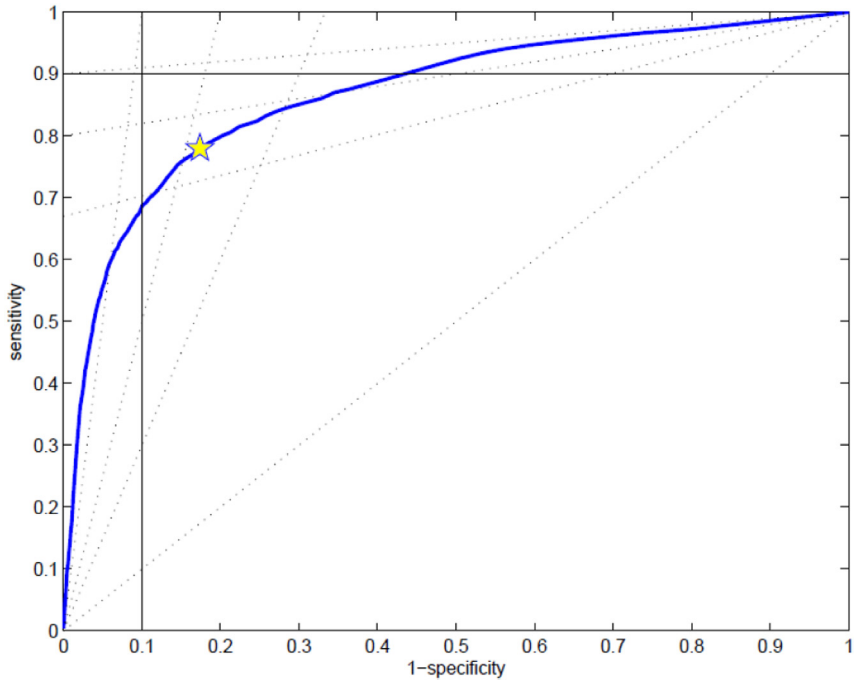

Fig. 4. ROC curve for .632 bootstrap estimates.

Table 7
632 bootstrap estimate LHR regions.

LHR+	3	5	≥ 10	LHR−	≤ 0.1	0.2	0.33
Cutoff	0.22	0.34	0.56–1	Cutoff	-	0–0.17	0.39

Fig. 4 shows the ROC curve for the .632 bootstrap estimate with AUC delivering moderate accuracy [41]. Clinically significant 90% sensitivity has a cutoff of 0.12 with 56% specificity. Similarly, 90% specificity is at cutoff 0.44 with 67% sensitivity.

Table 7 shows LHR+ regions with the potential to alter clinical decisions occur at cutoff values greater than 0.56, where cutoff values greater than 0.34 often provide useful information and values less than 0.22 rarely alter clinical decisions for positive results [19]. LHR− regions do not perform with the potential to alter clinical decisions. However, negative tests with cutoff values less than 0.17 often provide useful information and values over 0.39 for negative results rarely alter clinical decisions [19].

At the optimal cutoff value, the .632 bootstrap estimate performs at an indeterminate level, but very near *high accuracy*, often providing useful information, and clinical significance. The .632 bootstrap estimate LHR regions do perform within ranges of *providing useful information* and *potential to alter clinical decisions* for cases and controls. Thus, the overall performance of the classifier is useful for clinical decision making identify case and control hours in real time.

4. Discussion

4.1. Performance assessment

The best classifier correctly identifies 78% (69–94%) of severe sepsis and septic shock hours and 83% (76–94%) of SIRS and sepsis hours at the optimal cutoff value. However, sensitivity and specificity should ideally be over 90% to minimize false positives and false negatives [36]. The resubstitution best case estimate meets these criteria in Fig. 2. In Fig. 3 the posterior probability distribution for the worst case bootstrap estimate maintains a tail for controls. However, there is a more uniform distribution across case hours. Overall, variability observed over the cases reduces performance in this latter bootstrap estimate. The .632 bootstrap estimate splits the difference in these cases.

Among positive predictive results, severe sepsis and septic shock hours are correctly identified 10-36% of the time, which is low sensitivity. SIRS and sepsis hours, or negative results for severe sepsis and septic shock, are correctly identified 99-100% of the time. However, as noted, predictive values are biased by disease prevalence [38]. Thus, predictive values alone do not represent these test's inherent accuracy.

Clinically, a clinician needs to make an inference about the presence or absence of infection. In this clinical scenario, the likelihood ratios used here are better able to assess predictive test properties because they are independent of disease prevalence [9,20,34]. At the optimal cutoff values LHR+ does not include 1, instead a positive test result is obtained approximately 4.5 times (2.9–15.7) more often from a patient with severe sepsis or septic shock than from a patient with SIRS or sepsis. Similarly, LHR− does not include 1, and the posterior probability of obtaining a negative test result is less than twenty-seven-hundredths (range: 6/100 to 40/100) as likely in a patient with severe sepsis or septic shock than a patient with SIRS or sepsis. Thus, overall, both LHRs perform above the level of *rarely altering clinical decisions*, but do not yet reach the level of *often provides useful information* [19]. The overlap of posterior probabilities between cases and controls in Figs. 2–4 brings down the minimum, and thus overall, performance.

LHRs and MLRs provide more important information beyond a binary negative or positive value, by delineating which levels of test results yield clinically important information and which do not [19]. For the resubstitution estimate, the optimal cutoff value of 0.35 is sufficient as a test positive and test negative threshold (Table 3), where LHR+ > 0.3 and LHR− < 0.38 both offer potential to alter clinical decisions [19]. For the bootstrap estimate, the optimal cutoff of 0.30 represents the minimum cutoff value of clinical utility for LHR+ results (Table 5), but is within a region of no clinical utility for LHR− results. For the .632 bootstrap estimate (Table 7), the optimal cutoff value of 0.31 is the threshold for LHR+

results to provide useful information in clinical decision making and the LHR– minimum boundary for clinical utility.

When the bootstrap estimate posterior probabilities for cases has a more uniform distribution (Fig. 3) compared to the resubstitution estimate (Fig. 2), LHR– suffers, not LHR+. Hence, LHR+ remains useful for clinical decisions because a small overlap of posterior probabilities at higher values remains. LHR– becomes less useful for clinical decisions due to the greater overlap of cases and controls at lower probability values. Thus, for all three different estimates (resubstitution, bootstrap, .632 bootstrap), LHR+ provides useful diagnosis of sepsis cases, and the classifier is better at providing useful diagnostic information for cases than controls. This outcome suggests there remain useful physiological values able to distinguish control hours (N) from becoming classified as case hours (S) in the clinical data used.

MLR results (Tables 3 and 5) support the LHR results. Specifically, positive results may provide useful information, with increasing utility at greater probability values. However, negative results may not have the same utility. Importantly, increasing accuracy with greater probability values is independent of sepsis severity (Figs. 2 and 3).

AUC = 0.87 (0.78–0.99) values show moderate to high accuracy [41] and generally very good discriminative properties [13]. The 90% sensitivity and specificity points on the ROC curve are shown in Figs. 2–4, where the cutoff for sensitivity and specificity of 90% vary from the optimal value for each differet estimate. In general, clinically acceptable tradeoffs exist in using these values for the resubstitution estimate, but decline with significant reductions in specificity and sensitivity for the other two estimates.

AUC is the single measure summarising test peformance across all cutoff values, and is independent of prevalence [13]. It also enables statistical comparison of diagnostic tests [16,29], particularly for the same patient population or the same diagnostic question. Similarly, the diagnostic odds ratio (DOR) is an alternative way to compare tests. As potentially useful tests have DOR over 20, the KDE classifier had DOR = 16.83 (7.04–262), and thus may be potentially useful towards its maximum performance since these values span that threshold level.

Recently, a PhysioNet Challenge [45] was undertaken on early detection of sepsis from clinical data, although it occurred after submission of this research. It utilized a much larger patient data set than this research with 60,000 h of data and 40 clinical variables per hour. The results are not yet released to compare accuracy or prediction. In addition, the approach here, as noted, is time independent, whereas this challenge used a time dependent utility metric in lieu of the more statistically rigorous and clinically accepted likelihood ratios (LHR+ and LHR– values) and ROC curves used here. As noted, this work is time independent, and thus lower in burden in clinical variables and computation, but may lack the resolution of a time dependent metric capturing the evolution of sepsis in these, and a further added, clinical variables. It is a limitation of this work to be addressed.

4.2. Methodology

The classifier chosen used the PK (product-kernel), a 0.5 prior probability, and Cholesky transformation, on place 3 of the kernel product (KP), disease prevalence prior probability, and PCA transformation. KP assumes mutual independence among the rotated data components, but PK does not, holding a weaker assumption of independent components for the kernel. Hence, the clinical predictors are not assumed to be mutually independent, matching known data for sepsis and sepsis physiology.

The prior probability in Eq. (1) is essentially a scaling factor. When set equal to disease prevalence of 3.5% for sepsis cases ($\hat{\pi}_S = 0.035$ and thus $\hat{\pi}_N = 0.965$), the control term for non-sepsis hours in Eq. (1) becomes so large the optimal cutoff value is unreasonably small. Hence, prior probabilities were assumed to be chance, and set to 0.5, thus cancelling these terms in Eq. (1).

4.3. Clinical significance

The KDE classifier presented was designed for real-time diagnosis of severe sepsis and septic shock cases (S), as discriminated from SIRS and sepsis controls (N) using limited bedside data and a proven model-based metabolic marker of patient condition. The design presented used controls representing patient hours which were at risk of becoming a sepsis case (S). It is important to note, controls have a lower, less clinically urgent disease severity (SIRS, mild sepsis), and the discrimination needed is precise.

Thus, because the difference between cases and controls is smaller compared to classifying severe sepsis cases versus non-diseased control cases, there is a correspondingly lower statistical power to detect an effect. Hence, this approach in this work comprises a much stricter, rigorous, and clinically realistic test of the classifier. This choice of approach is justified considering the classifier is trying to discriminate the need for aggressive treatment. Finally, these issues also summarize the primary difficulty in classifying sepsis for diagnosis, where, in fact, the disease exists along a continuous spectrum despite specific disease level scores, and the resulting diagnostic classification is explicit and binary. Transforming a continuum into a discrete variable with classification will thus carry the risk of error, particularly when the difference is, as noted, narrow between the two sets being discrminated.

Overall, the performance of the classifier presented may be useful for real-time, non-invasive clinical decision making in sepsis patients. However, due to the narrow patient cohort used for validation, it is uncertain how it will perform when a broad spectrum of alternative diagnoses are available in the data and may show some cross over to these cases for misdiagnosis. Further research is thus required incorporating patients with non-infectious SIRS and non-infectious shock. It is hypothesized such cases would more likely change specificity, rather than sensitivity. Equally, such a broader test could improve results compraed to the conservative results presented, and a greater number of patients and cases could also improve performance.

Finally, this KDE classification model, as presented, is not time dependent. The probability density profiles in Eq. (1) are developed from hourly physiological data, but the time course is not considered from hour to hour. Hence, independence is assumed between physiological data and sepsis state hour to hour, with no influence from data or state in prior hours. There is thus room to improve the classifier by considering hourly sepsis transition probabilities given sepsis has a known general evolution over time.

5. Conclusions

Severe sepsis is a high morbidity, high mortality disease. Fast, accurate, real-time diagnosis would improve outcomes and cost, but do not yet exist. A classifier was designed to discriminate hourly physiological and model-based data comprising 213 h of severe sepsis and septic shock and 5858 h of SIRS and sepsis controls. The limited physiological data used is readily available at the bedside and already typically captured by computerized systems. Kernel density estimates of the Bayes classifier were implemented using product kernel estimates to create joint probability density profiles for severe sepsis cases versis controls for real-time diagnosis hour-to-hour.

The classifier performs with the greatest stability and accuracy when using the product kernel, 0.5 prior probabilities, and Cholesky transformation. The classifier shows good diagnostic capability. More specifically, compared to standard guidelines, it of-

ten provides useful additional information for clinical decision making, as well as increased accuracy with greater posterior probabilities and is independent from sepsis severity. Thus, the classifier provides a non-invasive, continuous estimate of whether or not severe sepsis or septic shock state exists in real time.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cmpb.2019.105295.

References

[1] D.C. Angus, W.T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, M.R. Pinsky, Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care, Crit. Care Med. 29 (7) (2001) 1303–1310.

[2] A. Blakemore, S.-H. Wang, A.L. Compte, G.M. Shaw, X.-W. Wong, J. Lin, T. Lotz, C.E. Hann, J. Geoffrey Chase, Model-based insulin sensitivity as a sepsis diagnostic in critical care, J. Diab. Sci. Technol. 2 (3) (2008) 468–477.

[3] R.C. Bone, R.A. Balk, F.B. Cerra, R.P. Dellinger, A.M. Fein, W.A. Knaus, R.M. Schein, W.J. Sibbald, Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. the ACCP/SCCM consensus conference committee. American college of chest physicians/society of critical care medicine, Chest J. 101 (6) (1992) 1644–1655.

[4] C. Chambrier, M. Laville, K Rhzioual Berrada, M. Odeon, P. Bouletreau, M. Beylot, Insulin sensitivity of glucose and fat metabolism in severe sepsis, Clin. Sci. 99 (4) (2000) 321–328.

[5] J. Geoffrey Chase, C.E. Hann, M. Jackson, J. Lin, T. Lotz, X.-W. Wong, G.M. Shaw, Integral-based filtering of continuous glucose sensor measurements for glycaemic control in critical care, Comput. Methods Progr. Biomed. 82 (3) (2006) 238–247.

[6] J. Geoffrey Chase, G. Shaw, A.L. Compte, T. Lonergan, M. Willacy, X.-W. Wong, J. Lin, T. Lotz, D. Lee, C. Hann, Implementation and evaluation of the sprint protocol for tight glycaemic control in critically ill patients: a clinical practice change, Crit. Care 12 (2) (2008) R49.

[7] C.A. Cooley, S.N. MacEachern, Classification via kernel product estimators, Biometrika 85 (4) (1998) 823–833.

[8] P.D. Docherty, J.G. Chase, T. David, Characterisation of the iterative integral parameter identification method, Med. Biol. Eng. Comput. 50 (2) (2012) 127–134.

[9] B. Dujardin, J. Van denEnde, A. Van Gompel, J.-P. Unger, P. Van derStuyft, Likelihood ratios: a real improvement for clinical decision making? Eur. J. Epidemiol. 10 (1) (1994) 29–36.

[10] B. Efron, R. Tibshirani, Improvements on cross-validation: the 632+ bootstrap method, J. Amer. Stat. Assoc. 92 (438) (1997) 548–560.

[11] B. Efron, Estimating the error rate of a prediction rule: improvement on cross–validation, J. Am. Stat. Assoc. 78 (382) (1983) 316–331.

[12] A. Evans, G.M. Shaw, A.L. Compte, C.-S. Tan, L. Ward, J. Steel, C.G Pretty, L. Pfeifer, S. Penning, F. Suhaimi, et al., Pilot proof of concept clinical trials of stochastic targeted (star) glycemic control, Ann. Intensive Care 1 (1) (2011) 1–12.

[13] J.E. Fischer, L.M. Bachmann, R. Jaeschke, A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis, Intensive Care Med. 29 (7) (2003) 1043–1051.

[14] E.J. Giamarellos-Bourboulis, P. Giannopoulou, P. Grecka, D. Voros, K. Mandragos, H. Giamarellou, Should procalcitonin be introduced in the diagnostic criteria for the systemic inflammatory response syndrome and sepsis? J. Crit. Care 19 (3) (2004) 152–157.

[15] S. Gupta, A. Sakhuja, G. Kumar, E. McGrath, R.S. Nanchal, K.B. Kashani, Culture-negative severe sepsis: nationwide trends and outcomes, Chest 150 (6) (2016) 1251–1259.

[16] J.A. Hanley, B.J. McNeil, et al., A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology 148 (3) (1983) 839–843.

[17] C.E. Hann, J.G. Chase, J. Lin, T. Lotz, C.V. Doran, G.M. Shaw, Integral-based parameter identification for long-term dynamic verification of a glucose–insulin system model, Comput. Methods Progr. Biomed. 77 (3) (2005) 259–270.

[18] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, R. Tibshirani, The Elements of Statistical Learning, 2, Springer, 2009 No. 1.

[19] R. Jaeschke, G.H. Guyatt, D.L. Sackett, G. Guyatt, E. Bass, P. Brill-Edwards, G. Browman, D. Cook, M. Farkouh, H. Gerstein, et al., Users' guides to the medical literature: Iii. how to use an article about a diagnostic test B. What are the results and will they help me in caring for my patients? JAMA 271 (9) (1994) 703–707.

[20] R. Jaeschke, G. Guyatt, D.L. Sackett, E. Bass, P. Brill-Edwards, G. Browman, D. Cook, M. Farkouh, H. Gerstein, B. Haynes, et al., Users' guides to the medical literatureiii. how to use an article about a diagnostic test A. Are the results of the study valid? JAMA 271 (5) (1994) 389–391.

[21] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, Apache II: a severity of disease classification system, Crit. Care Med. 13 (10) (1985) 818–829.

[22] A. Kumar, D. Roberts, K.E. Wood, B. Light, J.E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, et al., Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock, Crit. Care Med. 34 (6) (2006) 1589–1596.

[23] L. Langouche, S. Vander Perre, P.J. Wouters, A. D'Hoore, T. Krarup Hansen, G. Van denBerghe, Effect of intensive insulin therapy on insulin sensitivity in the critically ill, J. Clin. Endocrinol. Metab. 92 (10) (2007) 3890–3897.

[24] M.M. Levy, M.P. Fink, J.C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S.M. Opal, J.-L. Vincent, G. Ramsay, 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference, Intensive Care Med. 29 (4) (2003) 530–538.

[25] J. Lin, J.D. Parente, J.G. Chase, G.M. Shaw, Amy J. Blakemore, A.J. LeCompte, C. Pretty, N.N. Razak, D.S. Lee, C.E. Hann, et al., Development of a model-based clinical sepsis biomarker for critically ill patients, Comput. Methods Progr. Biomed. 102 (2) (2011) 149–155.

[26] J. Lin, N.N. Razak, C.G. Pretty, A.L. Compte, P. Docherty, J.D. Parente, G.M. Shaw, C.E. Hann, J.G. Chase, A physiological intensive control insulin-nutrition-glucose (icing) model validated in critically ill patients, Comput. Methods Progr. Biomed. 102 (2) (2011) 192–205.

[27] P. Martínez-Camblor, Nonparametric cutoff point estimation for diagnostic decisions with weighted errors, Rev. Colomb. Estad. 34 (1) (2011) 133–146.

[28] G.S. Martin, D.M. Mannino, S. Eaton, M. Moss, The epidemiology of sepsis in the united states from 1979 through 2000, N. Engl. J. Med. 348 (16) (2003) 1546–1554.

[29] B.J. McNeil, J.A. Hanley, H.H. Funkenstein, J. Wallman, Paired receiver operating characteristic curves and the effect of history on radiographic interpretation. ct of the head as a case study, Radiology 149 (1) (1983) 75–77.

[30] L. Mica, E. Furrer, M. Keel, O. Trentz, Predictive ability of the iss, niss, and apache ii score for sirs and sepsis in polytrauma patients, Eur. J. Trauma Emerg. Surg. 38 (6) (2012) 665–671.

[31] K.T. Moorhead, D. Lee, J.G. Chase, A.R. Moot, K.M. Ledingham, J. Scotter, R.A. Allardyce, S.T. Senthilmohan, Z. Endre, Classifying algorithms for sift-ms technology and medical diagnosis, Comput. Methods Progr. Biomed. 89 (3) (2008) 226–238.

[32] S.L. Murphy, J. Xu, K.D. Kochanek, Deaths: final data for 2010, Natl. Vital Stat. Rep. 61 (4) (2013) 1–117.

[33] A. Nakamura, H. Wada, M. Ikejiri, T. Hatada, H. Sakurai, Y. Matsushima, J. Nishioka, K. Maruyama, S. Isaji, T. Takeda, et al., Efficacy of procalcitonin in the early diagnosis of bacterial infections in a critical care unit, Shock 31 (6) (2009) 587–592.

[34] S.G. Pauker, J.P. Kassirer, The threshold approach to clinical decision making, N. Engl. J. Med. 302 (20) (1980) 1109–1117.

[35] J. Phua, W.J. Ngerng, K.C. See, C.K. Tay, T. Kiong, Hui F. Lim, M.Y. Chew, H.S. Yip, A. Tan, H.J. Khalizah, et al., Characteristics and outcomes of culture-negative versus culture-positive severe sepsis, Crit. Care 17 (5) (2013) R202.

[36] C. Pierrakos, J.-L. Vincent, et al., Sepsis biomarkers: a review, Crit Care 14 (1) (2010) R15.

[37] W.F.W. Muhd Shukeri, M.B. Mat-Nor, U.K. Jamaludin, F. Suhaimi, N.N. Abd. Razak, A.M. Ralib, Levels and diagnostic value of model-based insulin sensitivity in sepsis: a preliminary study, Indian J. Crit. Care Med. 22 (6) (2018) 402.

[38] J.E. Smith, R.L. Winkler, D.G. Fryback, The first positive: computing positive predictive value at the extremes, Ann. Intern. Med. 132 (10) (2000) 804–809.

[39] K.W. Stewart, J.G. Chase, C.G. Pretty, G.M. Shaw, Nutrition delivery of a model-based ICU glycaemic control system, Ann. Intensive Care 8 (1) (2018) 4.

[40] K.W. Stewart, C.G. Pretty, H. Tomlinson, F.L. Thomas, J. Homlok, S.N. Noémi, A. Illyés, G.M. Shaw, B. Benyó, J.G. Chase, Safety, efficacy and clinical generalization of the star protocol: a retrospective analysis, Ann. Intensive Care 6 (1) (2016) 24.

[41] J.A. Swets, Measuring the accuracy of diagnostic systems, Science 240 (4857) (1988) 1285–1293.

[42] B.M.P. Tang, G.D. Eslick, J.C. Craig, A.S. McLean, Accuracy of procalcitonin for sepsis diagnosis in critically ill patients: systematic review and meta-analysis, Lancet Infect. Dis. 7 (3) (2007) 210–217.

[43] V. Uyttendaele, J.L. Dickson, G.M. Shaw, T. Desaive, J.G. Chase, Untangling glycaemia and mortality in critical care, Crit. Care 21 (1) (2017) 152.

[44] B. Uzzan, R. Cohen, P. Nicolas, M. Cucherat, G.-Y. Perret, Procalcitonin as a diagnostic test for sepsis in critically ill adults and after surgery or trauma: a systematic review and meta-analysis, Crit. Care Med. 34 (7) (2006) 1996–2003.

[45] M.A. Reyna, C.S. Josef, R. Jeter, S.P. Shashikumar, M.B. Westover, S. Nemati, G.D. Clifford, A. Sharma, Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019, Crit. Care Med. (2019) [Epub ahead of print], doi:10.1097/CCM.0000000000004145.