PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Development and internal validation of a simple prognostic score for early sepsis risk stratification in the emergency department
AUTHORS	Liu, Bofu; Li, Dongze; Cheng, Yisong; Yu, Jing; Jia, Yu; Zhang, Qin; Liu, Yanmei; Cao, Yu

VERSION 1 – REVIEW

REVIEWER	Patricia Stassen Maastricht University Medical Centre+, Internal Medicine
REVIEW RETURNED	13-Dec-2020

GENERAL COMMENTS	Dear editor, dear authors, This manuscript covers an important area of sepsis research, namely early identification of sepsis patients at high risk for adverse outcomes in the emergency department. While all recent scores have been developed from an ICU perspective and extrapolated to the ED/prehospital environment, this study focuses on recognition/identification of those at risk immediately after arrival in the ED. The study has included enough patients, with rather high mortality rates. Although many aspects are nicely addressed, I have some major concerns/statements/remarks. Methodology: I miss a statement on STARD guidelines for reporting and some items are lacking. for instance, how was the study sample assembled? Consecutive patients, convenience cohort? I miss a flow chart. Mortality was very high, I miss info on the number of patients who were admitted to ICU, with shock and on comorbidity. To be able to extrapolate the results of this study, the study population and setting should be made more clear. The score was validated in the validation group. I cannot find info on the number of events (mortality, ICU admissions) within this cohort. Were there enough events to be able to serve as a real validation cohort. I miss a sample calculation. Then, with regard to the analyses that were performed. Why did you check on the predictive value of lab items, while you were not planning to incorporate these into your model? It is a bit confusing for the reader -> leave these items out of your analyses. I miss statements on missing items-> how were these handled? You presented AUCs and compared these with??. In fig 2 p values are presented -> which comparisons do they refer to?

	Fig 1: the legends of the axes are missing I am not able to follow/comprehend the analysis on the NRI (Table 4). On the vertical axis, SOFA scores are mentioned and divided into 3 groups <=0.05 >0.2 and in between. What do these values represent (The SOFA score works with points and cut off value of 2)? The same is not clear on the horizontal axis regarding the SSEPS score. In short, this table and thereby the analysis on the NRI is incomprehensible to me. I miss info on the analsysis on IDI. In summary, it is hard to establish the quality of the analyses that were performed. The endpoint is not completely clear, probably because of a language issue: what is sepsis 28-day all cause mortality? Is it sepsis-related? or is it all-cause mortality? The same is true for the secondary outcome: "...including mechanical ventilation, and admission to ICU" -> the word including suggests that there were more items considered as secondary endpoint. The prevalence/incidence of these outcomes should be better depicted in the manuscript and probably incorporated in the tables/figs. Figure 3 is also not easy to follow, probably by mistakes made in the Figure legends, where I cannot find the SSEPS score. Then with regard to de assignment of points for the 4 items that were eventually incorporated in the model (Table 3). Why was a heartrate of 107-123 assigned with only 1 oint? It would be logical that with increasing heartrate, the points assigned woudl increase as well (doese-response curve) -> pleas elucidate. Wtih regard to age: why not choose points per year instead of divinding into 4 categories? What is the definition of coma? In concusion, probably, the methdology was okay, but improvements in the way the results/methods were presented is really necessary, both for the reader and to be able to really judgde the scientific quality of this study. Then, minor issues: regarding the abstract: I miss the word sepsis in the first sentence (sepsis patients, adults). Outcome mesaures: the analyses are not outcome measures. Results: I miss the number of included patients. I miss exact results on eg AUC etc. Regarding the rest of the manuscript: Introduction: is a bit too long. I miss short info on the study desgin. Methods: which test was used to compare aucs of the scores? Discussion: I miss the limitations. I question the remark on the qSOFA that it needs instrumental measurements, as only systolic BP, respiratory rate and consiousness are incorported, just as in the SEPSS score.

REVIEWER	Oliver Redfern Oxford University, Nuffield Department of Clinical Neurosciences
REVIEW RETURNED	18-Jan-2021

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript. This is a single centre retrospective study of patients who met sepsis

	criteria in the emergency department. The authors have developed and internally validated a novel score (SSEPS) to predict 28 day mortality. I am afraid I found the manuscript lacking key details in places, which made it difficult to assess some aspects of the study. I would recommend the authors report to the TRIPOD guidelines (https://www.equator-network.org/reporting-guidelines/tripod-statement/) when revising the manuscript. Specific comments: 1. The abstract does not clearly describe the primary/secondary outcomes or report key numerical results (e.g. AUC for SSEPS vs other scoring systems). In the objectives, I was unsure what the authors meant by the term “instrumental devices”. 2. Page 3/26;Lines 47-52: Could the authors clarify what they mean by “especial care”? 3. The introduction reviews limitations of existing scoring systems but could be much clearer in how this links to the study rationale. In particular, referring to the “out of hospital” setting could be confusing, as they then refer to “outside of the ICU or during ED admissions study”. Given the study is focused on ED attendance, could the authors clarify which setting (e.g. pre-hospital, ED or on the general wards) they are targeting? 4. Study population: The authors use a patient cohort described in reference 10. However, could they consider including a flowchart to the supplementary material? Did they only exclude patients who suffered “cardiac or respiratory arrest” pre-hospital or those who had a subsequent arrest? Could they clarify which “vasoactive drugs” met the exclusion criterion? 5. Section 2.3: I would suggest using the term “development set”, rather than “derivation”. 6. Could the authors explain what they mean by “Coma was identify that GCS is less than 13”? 7. Score development: Why were these specific laboratory tests chosen? Why did the authors not consider interactions in their model e.g. between urea nitrogen and eGFR? Could the authors provide more detail on how the references groups were created? Table 3 suggests the authors used a second logistic regression model to assign the score points – is this correct? 8. Score validation: Assessment of discrimination (AUC) and calibration should by presented on the validation set, particularly when comparing against other scoring systems. Figure 2 likely represents optimistic performance. Section 3.6: why were only SOFA and APACHE-II compared in the decision curve analysis? 9. Discussion: the authors do not clearly present the limitations of their work. For example, that the score was only internally validated on a relatively small cohort of patients on admission to ED.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Dr. Patricia Stassen, Maastricht University Medical Centre

Comments to the Author:

Dear editor, dear authors,

This manuscript covers an important area of sepsis research, namely early identification of sepsis patients at high risk for adverse outcomes in the emergency department. While all recent scores have been developed

from an ICU perspective and extrapolated to the ED/prehospital environment, this study focuses on recognition/identification of those at risk immediately after arrival in the ED.

The study has included enough patients, with rather high mortality rates.
Although many aspects are nicely addressed, I have some major concerns/statements/remarks.

Methodology:
1. I miss a statement on STARD guidelines for reporting and some items are lacking. for instance, how was the study sample assembled? Consecutive patients, convenience cohort? I miss a flow chart.

Thank you for your suggestion. We have described all the items on the TRIPOD checklist, and append the completed checklist as supporting information.

2. Mortality was very high, I miss info on the number of patients who were admitted to ICU, with shock and on comorbidity. To be able to extrapolate the results of this study, the study population and setting should be made more clear.

Thanks very much for your advice. We have added supplementary Table 2S to show the informations that adverse outcome of sepsis patients categorized by tertiles of simple sepsis early prognostic score (SSEPS: T1 (0–6 points) vs. T2 (7–9 points) vs. T3 (10–21 points)) among development cohort.

Variable	Simple sepsis early prognostic score			P
	T1（n=329）	T2（n=213）	T3（n=43）	
ICU admissions	97(29.5%)	112(52.6%)	33(76.7%)	<0.001
Mechanical ventilation	100(30.4%)	113(53.1%)	37(86.0%)	<0.001
Sepsis Shock	18(5.5%)	35(16.4%)	16(37.2%)	<0.001
Mortality	41(12.5%)	61(28.6%)	23(53.5%)	<0.001

3. The score was validated in the validation group. I cannot find info on the number of events (mortality, ICU admissions) within this cohort. Were there enough events to be able to serve as a real validation cohort. I miss a sample calculation.

（1） Thanks very much for your advice. We have added supplementary Table 3S to show the informations that adverse outcome of sepsis patients categorized by simple sepsis early prognostic score among validation cohort

Variable	Simple sepsis early prognostic score			P
	Q1（n=79）	Q2（n=79）	Q3（n=78）	
ICU admissions	18(22.8%)	26(32.9%)	47(60.3%)	<0.001
Mechanical ventilation	20(25.3%)	23(29.1%)	53(67.9%)	<0.001
Sepsis Shock	5(6.3%)	9(11.4%)	25(32.1%)	<0.001
Mortality	3(3.8%)	11(13.9%)	34(43.6%)	<0.001

（2） Thanks very much for your advice. We have added sample calculation method in the manuscript methods.

Based on the development cohort data, the mortality rate is approximately 20%. To satisfy this difference with 80% power at 5% significance (2-tailed), assuming that the predicted value of the scoring system for death is higher than the 0.75 AUC value, at least 24 deaths are required, and the number of validation group population was 120.

4. Then, with regard to the analyses that were performed. Why did you check on the predictive value of lab items, while you were not planning to incorporate these into your model? It is a bit confusing for the reader -> leave these items out of your analyses.

Thanks for your question. In order to fully correct the model, all the significant variables in single factor results were included in the regression model. We strongly agree with the good opinion of reviewer, and we realized it was a wrong and a confusing work. In deed, our purpose is to find the risk factors that are not rely on laboratory indicators and to build a evaluation model of these risk factors. So, we have rebuild the Logistic regression model as the follow.

Variables	**Univariate analysis**			**Multivariate analysis**		
	OR	95%CI	P	OR	95%CI	P
Age	1.020	1.000-1.040	0.052	1.032	1.008-1.056	0.010
Heart rate	1.009	0.999-1.018	0.081	1.055	1.010-1.100	0.040
Respiratory rate	1.052	1.017-1.088	0.003	1.042	1.003-1.083	0.034
Altered consciousness	3.721	2.385-5.806	＜0.001	3.606	2.296-5.664	<0.001

SBP, mmHg	0.999	0.991-1.007	0.807	1.003	0.989- 1.016	0.693
DBP, mmHg	0.994	0.982-1.006	0.307	0.987	0.969-1.006	0.175
SPO_2, %	0.979	0.955-1.004	0.099	0.991	0.963-1.020	0.530
Temperature, ℃	0.902	0.783-1.039	0.151	0.898	0.771-1.047	0.169

SBP: systolic blood pressure; DBP: diastolic blood pressure; SPO_2: oxygen saturation.

5. I miss statements on missing items-> how were these handled?

Thanks very much for your advice. It is hard to miss the key data in our statistics. In the case of missing data, continuous variable was filled with median and we treated it as negative if the categorical variables missed. In the case of missing more than 5%, this variable was directly excluded. We have added a statement about missing items handling in our manuscript section 2.6

6. You presented AUCs and compared these with??. In fig 2 p values are presented -> which comparisons do they refer to?

Thanks for your review. Sorry for the comparison is really not fully expressed in Fig 2, I have modified and supplemented Fig 2. The SSEPS take as a reference in the ROC curve analysis.

7. Fig 1: the legends of the axes are missing
Thanks for your careful review. Sorry for missing the axes, I have add the legends of the axes in the Fig 1.

8. I am not able to follow/comprehend the analysis on the NRI (Table 4). On the vertical axis, SOFA scores are mentioned and divided into 3 groups <=0.05 >0.2 and in between. What do these values represent (The SOFA score works with points and cut off value of 2)? The same is not clear on the horizontal axis regarding the SSEPS score. In short, this table and thereby the analysis on the NRI is incomprehensible to me. I miss info on the analsysis on IDI. In sum

ary, it is hard to establish the quality of the analyses that were performed.

Thanks very much for your advice. The Statistical analysis indeed neglected the important information of NRI cutoff points. The analysis of IDI represents the comparision in discrimination of the scores as a continuous variable, the meaning of this value is similar to that of differences in AUC curve value, but the algorithm of IDI is different from that of ROC. We added references to the *Statistical Analysis* section.

The following is our additional description in the Statistical analysis.

we take 5% as a cut-off value of low risk of mortality. Considering that mortality about 20% in our study, thus, in the NRI analysis, patients with risk of mortality less than 5%, 5% to 20%, and more than 20% were divided into low-risk, moderate, and high risk groups, respectively. (Refer to the previous literature reports)[***Widera C, European heart journal, 2012, 33(9): 1095-1104], [Costa F, The Lancet, 2017, 389(10073): 1025-1034].***

9. The endpoint is not completely clear, probably because of a language issue: what is sepsis 28-day all cause mortality? Is it sepsis-related? or is it all-cause mortality? The same is true for the secondary outcome: "...including mechanical ventilation, and admission to ICU" -> the word including suggests that there were more items considered as secondary endpoint. The prevalence/incidence of these outcomes should be better depicted in the manuscript and probably incorporated in the tables/figs. Thanks for your question, our primary endpoint was 28-day all-cause mortality. Moreover, We mistakenly identified mechanical ventilation and admission to intensive care unit as secondary endpoints, in fact, these adverse events was used to indirectly evaluate the effectiveness of SSEPS. Thank you for helping me correct this problem. I have deleted this content in the original manuscript.

11. Figure 3 is also not easy to follow, probably by mistakes made in the Figure legends, where I cannot find the SSEPS score.

Thanks for your careful review. SSEPS score was expressed as solid lines in the graph, and they have a higher predictive value than other scores under any of the same thresholds. [***Yamamoto S, BMJ open, 2015, 5(4)]***

12. Then with regard to de assignment of points for the 4 items that were eventually incorporated in the model (Table 3). Why was a heartrate of 107-123 assigned with only 1 point? It would be logical that with increasing heartrate, the points assigned woudl increase as well (doese-response curve) -> pleas elucidate. Wtih regard to age: why not choose points per year instead of divinding into 4 categories? What is the definition of coma?

Thanks for your careful review. After careful check, we found that in our table-making process, we mistakenly took the heart rate of 95-107 β value and the assigned score as the heart rate of 102-123 relevant value. In the article, we have made changes. Thank you very much for your correction, and we have also modified it.

If we take continuous ages into the regression model to analysis, we can calculated the increased risk for each additional year of patients. But since there was no baseline for comparison, it

was impossible to calculate age-specific scores for each patient.***[Song C, Scientific reports, 2018, 8(1): 1-8]***

We didn't think coma was a good word for a state of consciousness, so we changed it to altered mentation（Moderate/ Severe）. Thank you very much for pointing out that we are not suitable expression for a state of consciousness. Because it's not always easy to recognize at 13 to 15 points for the state of consciousness changes in clinical. We defined GCS＜13 as altered consciousness. ***[McKee A C, Handbook of clinical neurology, 2015, 127: 45-66]***

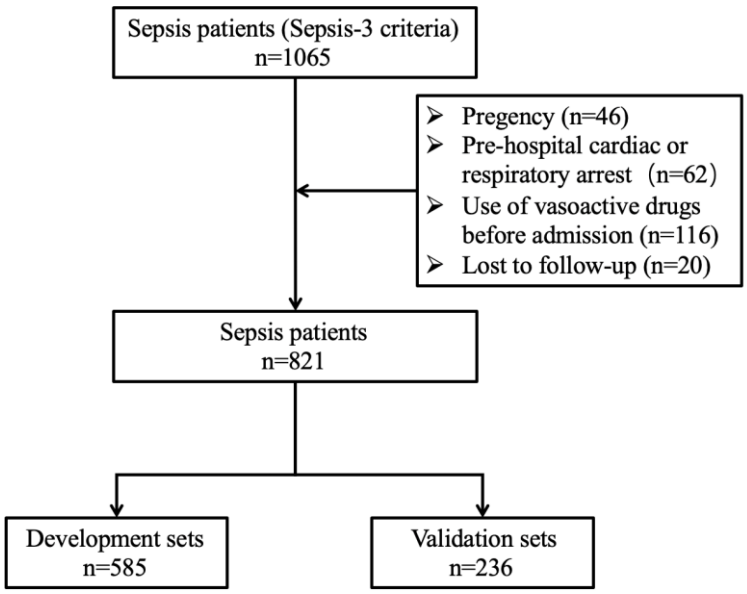
Then, minor issues:
regarding the abstract: I miss the word sepsis in the first sentence (sepsis patients, adults).

*T*hanks for your careful review. We have revised this sentence.

Outcome mesaures: the analyses are not outcome measures.

Thanks for your careful review. We have revised outcome measures.

Results: I miss the number of included patients. I miss exact results on eg AUC etc.
Thanks for your careful review. We have informations about the number of included patients is 821.We have added exact results on The AUCs of SSEPS for predicting sepsis 28-day mechanical ventilation (AUC:0.718, 95%CI: 0.650–0.785) and 28-day admission to intensive care unit (AUC:0.710, 95%CI: 0.632–0788).



Regarding the rest of the manuscript:
Introduction: is a bit too long. I miss short info on the study desgin.
Thanks for your comments. We have condensed our introduction to highlight the weaknesses of existing scoring systems and what our scoring system can do to make up for these deficiencies

Methods: which test was used to compare aucs of the scores?

Thanks very much for your advice. We used the Delong test to compare AUCs of the scores. We've described it in the statistics section. ***[Demler O V, Statistics in medicine, 2012, 31(23): 2577-2587.]***

Discussion:
I miss the limitations.

Thanks very much for your advice. We have present our work limitations in the discussion.

I question the remark on the qSOFA that it needs instrumental measurements, as only systolic BP, respiratory rate and consiousness are incorported, just as in the SEPSS score.

Thanks for your careful review. Thank you very much for pointing out this important error to us. In the description of qSOFA inadequacies, we mainly focused on its poor validity and discriminating ability. In addition, In our regression model, we included readily available non-laboratory indicators(e.g. age, heart rate, respiration rate, altered consciousness, SBP, DBP, SPO_2 and Temperature) that were not considered in the qSOFA during early phase ED admission or pre-hospital. However, after regression verification, only age, heart rate, respiration rate and altered consciousness were likely to be independent predictors of prognosis, So we updated our sepsis assessment tool based on the shortcomings of qSOFA. Finally, in our data results, it is found that qSOFA does have poor validity and discriminating ability for sepsis prognosis assessment. We have deleted this incorrect description in our manuscript.

Reviewer: 2

Dr. Oliver Redfern, Oxford University
Comments to the Author:
Thank you for the opportunity to review this manuscript.

This is a single centre retrospective study of patients who met sepsis criteria in the emergency department. The authors have developed and internally validated a novel score (SSEPS) to predict 28

day mortality.

I am afraid I found the manuscript lacking key details in places, which made it difficult to assess some aspects of the study. I would recommend the authors report to the TRIPOD guidelines (https://www.equator-network.org/reporting-guidelines/tripod-statement/) when revising the manuscript.

Thank you for your recommandation. We have reported to the TRIPOD guidelines and append the completed checklist as supporting information.

Specific comments:
1. The abstract does not clearly describe the primary/secondary outcomes or report key numerical results (e.g. AUC for SSEPS vs other scoring systems). In the objectives, I was unsure what the authors meant by the term “instrumental devices”.

Thank you for your suggestions. First, We have revised the description of primary/secondary outcomes in the manuscript. Second, I'm sorry about the lack of explanation to instrumental devices, In fact, instrumental devices mainly refers to laboratory tests, so I modified the expression in the manuscript.
2. Page 3/26;Lines 47-52: Could the authors clarify what they mean by “especial care”?

Thanks for your careful review. Nursing care is one of the important factors to improve the prognosis of sepsis patients. But this article does not focus on this aspect, in order to avoid readers' misunderstanding, we have deleted it.
3. The introduction reviews limitations of existing scoring systems but could be much clearer in how this links to the study rationale. In particular, referring to the “out of hospital” setting could be confusing, as they then refer to “outside of the ICU or during ED admissions study”. Given the study is focused on ED attendance, could the authors clarify which setting (e.g. pre-hospital, ED or on the general wards) they are targeting?

Thanks for your comments. The main shortcoming of existing scoring systems is that most of them require laboratory tests and cannot meet the requirements of early admission and prehospital prognostic assessment. vital signs of patients which reflect the host response during the early sepsis stages are usually obtained earlier than other information. Therefore, We aimed at constructing and validating an simple sepsis early prognostic score (SSEPS) based on vital signs that could predict the prognosis and severity of sepsis patients in the early stage of the emergency department admission and even pre-hospital. We realized that our description in this section is very inappropriate and we have revised the confusing expression.

4. Study population: The authors use a patient cohort described in reference 10. However, could they consider including a flowchart to the supplementary material? Did they only exclude patients who

suffered “cardiac or respiratory arrest” pre-hospital or those who had a subsequent arrest? Could they clarify which “vasoactive drugs” met the exclusion criterion?

Thanks for your suggestions. we have added a flowchart to the supplementary material. I am sorry to made a confusing expression, we excluded the patients with pre-hospital cardiac or respiratory arrest and the use of vasoactive drugs, mainly because drug interventions might alter the physiological parameters of patients, which would affect the establishment and validation of our model. We have revised it in this manuscript.

5. Section 2.3: I would suggest using the term “development set”, rather than “derivation”.

Thanks for your careful review and valuable suggestions. We have modified it in the manuscript.

6. Could the authors explain what they mean by “Coma was identify that GCS is less than 13”?

Thanks for your question. We didn't think coma was a good word for a state of consciousness, so we changed it to altered mentation（Moderate/ Severe）. Thank you very much for pointing out that we are not suitable expression for a state of consciousness. Because it's not always easy to recognize at 13 to 15 points for the state of consciousness changes in clinical. We defined GCS＜13 as altered mentation. ***[McKee A C, Handbook of clinical neurology, 2015, 127: 45-6***

7. Score development: Why were these specific laboratory tests chosen? Why did the authors not consider interactions in their model e.g. between urea nitrogen and eGFR?

Thanks for your question. These laboratory tests have been shown to be risk factors in previous studies, so these may be potential confounding factors. We strongly agree with the good opinion of reviewers. Now we realize that our purpose is to find the risk factors that are not laboratory indicators and build the evaluation model of these risk factors, so we rebuild the regression model.

Variables	**Univariate analysis**			**Multivariate analysis**		
	OR	95%CI	P	OR	95%CI	P
Age	1.020	1.000-1.040	0.052	1.032	1.008-1.056	0.010
Heart rate	1.009	0.999-1.018	0.081	1.055	1.010-1.100	0.040
Respiratory rate	1.052	1.017-1.088	0.003	1.042	1.003-1.083	0.034
Altered consciousness	3.721	2.385-5.806	＜0.001	3.606	2.296-5.664	<0.001
SBP, mmHg	0.999	0.991-1.007	0.807	1.003	0.989- 1.016	0.693

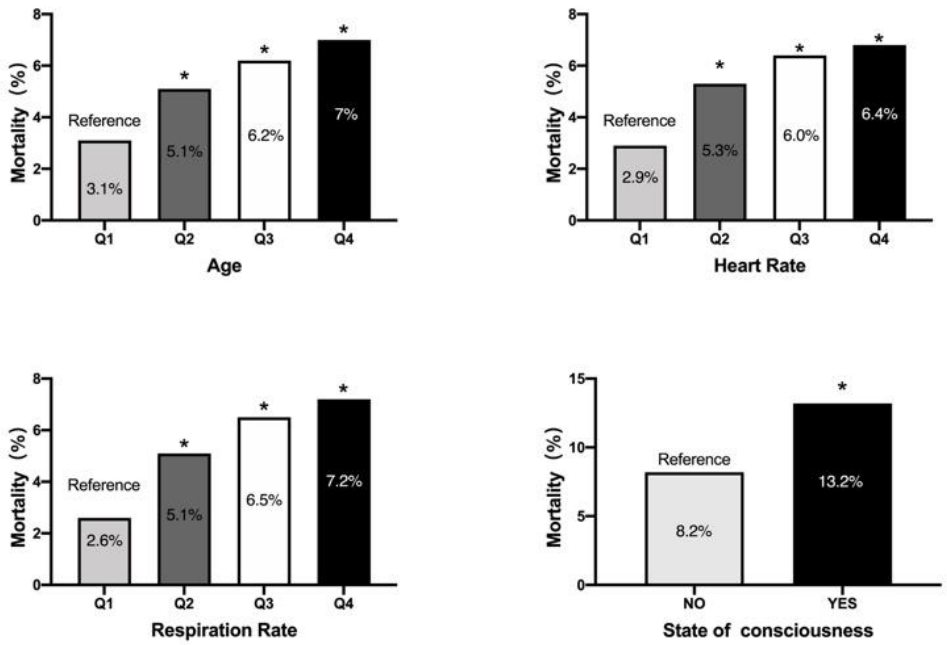
DBP, mmHg	0.994	0.982-1.006	0.307	0.987	0.969-1.006	0.175
SPO_2, %	0.979	0.955-1.004	0.099	0.991	0.963-1.020	0.530
Temperature, ℃	0.902	0.783-1.039	0.151	0.898	0.771-1.047	0.169

SBP: systolic blood pressure; DBP: diastolic blood pressure; SPO_2: oxygen saturation.

8. Could the authors provide more detail on how the references groups were created? Table 3 suggests the authors used a second logistic regression model to assign the score points – is this correct?

Thanks very much for your advice. Sorry for not adequately expressing how we created the references groups. We selected the lowest risk group like in our supplementary figure 2S as the reference group to ensure that each subgroup had a score greater than 0.

Because other confounding factors will affect the weight ratio of these risk factors, in addition, the variable in the original model is a continuous variable. Therefore, categorical variables are used to reconstruct the score to achieve the process of assigning scores to different groups. This method of assigning points has been well-recognized in previous published articles. ***[Song C, Scientific reports, 2018, 8(1): 1-8]***



9. Score validation: Assessment of discrimination (AUC) and calibration should by presented on the validation set, particularly when comparing against other scoring systems. Figure 2 likely represents

optimistic performance. Section 3.6: why were only SOFA and APACHE-II compared in the decision curve analysis?

Thanks very much for your advice. We have added the informations about the AUC comparison between SSEPS and other scores in the validation group and append the supplementary Figure 3s as supporting information. In Validation Group, the AUCs of SSEPS, SOFA and APACHE-II were the highest of all the scoring systems, and there is no statistical difference between them. So we further comparing these scores using decision curve analysis. Thank you for reminding us to make up for this logical breakpoint.

10. Discussion: the authors do not clearly present the limitations of their work. For example, that the score was only internally validated on a relatively small cohort of patients on admission to ED .

Thanks very much for your advice. We have present our work limitations in discussion.

VERSION 2 – REVIEW

REVIEWER	Oliver Redfern Oxford University, Nuffield Department of Clinical Neurosciences
REVIEW RETURNED	11-Apr-2021

GENERAL COMMENTS	General comments: I thank the authors for their comprehensive response to previous comments and for their substantial efforts to revise this manuscript. However, I feel some details in the manuscript remain unclear. Given that the relationship between SSEPS and additional outcomes (ICU admission, mechanical ventilation, and shock) feature prominently in the results and supplementary material, I would suggest that it would be clearer to the reader if they were described as secondary outcomes throughout the manuscript. Specific comments: 1. Title: I suggest the authors revise the title to avoid repetition and the ambiguity of the word “early”. My suggestion would be: Development and internal validation of a simple prognostic score for early sepsis risk stratification in the emergency department 2. Abstract: Key numerical results (e.g. AUC) are still missing. 3. Introduction: The introduction is much clearer but repeatedly refers to “dynamic risk assessment”. This was not tested in the study, which only used the first set of vital signs taken in the ED. 4. Methods: a. ICU admission, mechanical ventilation and shock as outcomes should be defined/mentioned in the methods, as these are reported in model validation. b. “In the case of missed value more than 5%, this variable was removed directly” -- could the authors please clarify this sentence? Did this affect any vital sign measurements? c. “We chose the resulting independent laboratory examination risk predictors with $P < 0.05$ to set up the SSEPS scoring system” -- this sentence could imply laboratory tests were used in SSEPS. I

	suggest instead explicitly listing the predictors used to develop the model (e.g. vital signs and age). d. AUC is a measure of discrimination, not “clinical utility” 5. Results: a. In Section 3.2 the authors say that GCS was higher in patients who died, Table 1 states the opposite. b. In Section 3.3, what do the authors mean by “After adjusting for relative risk factors”? c. In Sections 3.5/3.6, reporting of AUCs, confidence intervals and p-values is inconsistent. I suggest reporting both confidence intervals and p-values for all scores/comparisons. The authors state in their response that the Delong test was used to compare AUCs relative to SSEPS – could they make this clearer to the reader? d. Calibration/discrimination figures on the development set are presented in the main manuscript, whereas equivalent results on the validation set are in the supplementary material -- I suggest reversing. 6. Discussion: a. “After adjusting for confounding factors, an elevated SSEPS risk score was an independent predictor of adverse outcomes, irrespective of the severity of sepsis” -- could the authors please clarify which results they are referring to?

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

Dr. Oliver Redfern, Oxford University

Comments to the Author:

General comments:

I thank the authors for their comprehensive response to previous comments and for their substantial efforts to revise this manuscript. However, I feel some details in the manuscript remain unclear.

Given that the relationship between SSEPS and additional outcomes (ICU admission, mechanical ventilation, and shock) feature prominently in the results and supplementary material, I would suggest that it would be clearer to the reader if they were described as secondary outcomes throughout the manuscript.

Thank you very much for your recognition of our work and your advise. The events of ICU admission, mechanical ventilation and shock have been described as secondary endpoints in our manuscript, and we modified the relevant content throughout the manuscript.

Specific comments:

1.Title: I suggest the authors revise the title to avoid repetition and the ambiguity of the word “early”. My suggestion would be: Development and internal validation of a simple prognostic score for early sepsis risk stratification in the emergency department

Thank you for your advise. We have revised the title of the manuscript to avoid repetition and the ambiguity of the word “early”.

“Development and internal validation of a simple prognostic score for early sepsis risk stratification in the emergency department”

2.Abstract: Key numerical results (e.g. AUC) are still missing.

Thanks for your suggestion. We have added the key numerical results in the abstract results section.

3.Introduction: The introduction is much clearer but repeatedly refers to “dynamic risk assessment”. This was not tested in the study, which only used the first set of vital signs taken in the ED.

Thanks very much for your advice. We have sufficiently recognized the deficiencies and remove “dynamic risk assessment”

4.Methods:

a.ICU admission, mechanical ventilation and shock as outcomes should be defined/mentioned in the methods, as these are reported in model validation.

Thanks for your suggestion. We defined/mentioned the ICU admission, mechanical ventilation and shock as secondary endpoints.

“Our primary endpoint was 28-day all-cause mortality, and the secondary endpoint was the incidence of ICU admission, mechanical ventilation and shock. ICU admission was performed when severe dystrophia, severe hydroelectrolyte balance and unstable vital signs occurred, and the organ supports were needed. Mechanical ventilation was defined as patients who require non-invasive ventilation or invasive ventilation due to hypoxic respiratory failure. Shock was defined as patients with hypotension that persists despite adequate fluid resuscitation and requires vasopressors to support perfusion.”

b.“In the case of missed value more than 5%, this variable was removed from baseline table directly” -- could the authors please clarify this sentence? Did this affect any vital sign measurements?

Thanks for your good question. Vital signs are the basic information of each admitted patient, we can guarantee the integrity of this part of the data. And our missing data is mainly the part of laboratory inspection results. Therefore, these missing data have little effect on our research results.

“We used the median to replace the laboratory variables missing less than 5% and if missing variables greater than 5% were not analyzed in the baseline.”

c.“We chose the resulting independent laboratory examination risk predictors with $P < 0.05$ to set up the SSEPS scoring system” -- this sentence could imply laboratory tests were used in SSEPS. I suggest instead explicitly listing the predictors used to develop the model (e.g. vital signs and age).

Thanks for your suggestion, your suggestion can reduce the misunderstanding of readers. We have explicitly listed the predictors with $P < 0.05$ to set up the SSEPS scoring system.

“We chose the age, heart rate, respiratory rate, altered consciousness, systolic blood pressure(SBP), diastolic blood pressure(DBP), oxygen saturation(SPO2) and temperature with $P < 0.05$ to set up the SSEPS scoring system”

d.AUC is a measure of discrimination, not “clinical utility”

Thank you for correcting our incorrect wording. “clinical utility” was replaced by “discrimination”

5.Results:

a.In Section 3.2 the authors say that GCS was higher in patients who died, Table 1 states the opposite.

Thanks for your careful review. In the clinical practice, lower GCS scores predict worse neurological outcome. After careful check, patients who died had lower GCS. We revised this mistake.

b.In Section 3.3, what do the authors mean by “After adjusting for relative risk factors”?

Thanks for your careful review. We realized that our expression was confusing, so we deleted confusing words.

c.In Sections 3.5/3.6, reporting of AUCs, confidence intervals and p-values is inconsistent. I suggest reporting both confidence intervals and p-values for all scores/comparisons. The authors state in their response that the Delong test was used to compare AUCs relative to SSEPS – could they make this clearer to the reader?

Thanks for your suggestion. We have verified the results carefully and reported both confidence intervals and p-values for all scores/comparisons in Sections 3.5 and 3.6 to

ensure consistency.

d.Calibration/discrimination figures on the development set are presented in the main manuscript, whereas equivalent results on the validation set are in the supplementary material -- I suggest reversing.

Thanks for your constructive suggestion, which will be very significant for our future. We have reversed the order to show the Calibration/discrimination figures for development set and validation set.

6.Discussion:

a.“After adjusting for confounding factors, an elevated SSEPS risk score was an independent predictor of adverse outcomes, irrespective of the severity of sepsis” -- could the authors please clarify which results they are referring to?

Thanks for your suggestion, We have realized that our statement was inaccurate, so we have revised it.

“Our data shows that patients with higher SSEPS scores had higher risks of 28-day all-cause mortality and incidence of ICU admission, mechanical ventilation and shock. Irrespective of the severity of sepsis, the prognostic value of the SSEPS score was superior to that of qSOFA, MEWS, MEDS SOFA and APACHE II”

VERSION 3 – REVIEW

REVIEWER	Oliver Redfern Oxford University, Nuffield Department of Clinical Neurosciences

REVIEW RETURNED	01-Jun-2021

GENERAL COMMENTS	I thank the authors again for their efforts to revise this manuscript. Overall, the clarity of the manuscript has improved significantly since initial submission. I remain unsure whether the superior performance of SSEPS over the other scores will remain after external validation, given the small sample size. Please accept my apologies if some of my previous comments were unclear, but I would suggest some minor adjustments to the following responses. • Missing data: The authors say in their response "Vital signs are the basic information of each admitted patient, we can guarantee the integrity of this part of the data." Does this mean no vital signs were missing? If so, please state this in the methods. For laboratory tests, the authors say they used median imputation, which is reasonable, although multiple imputation would be preferable. However, of the scores tested, only MEDS/SOFA/APACHE-II use lab data. Again, I would state this in the methods for clarity if this was the rationale for imputation, otherwise no imputation would be required for development/validation of SSEPS. • Calibration: Calibration on the development set is reported in the abstract and Figure 1. As per my previous comment, calibration on the validation set should be presented. • Strength/limitations bullet points: Although mentioned in the discussion, the authors do not include the small sample size (236 patients in the validation set) here.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 2

Dr. Oliver Redfern

Oxford University

Comments to the Author:

I thank the authors again for their efforts to revise this manuscript. Overall, the clarity of the manuscript has improved significantly since initial submission. I remain unsure whether the superior performance of SSEPS over the other scores will remain after external validation, given the small sample size.

Please accept my apologies if some of my previous comments were unclear, but I would suggest some minor adjustments to the following responses.

• Missing data: The authors say in their response “Vital signs are the basic information of each admitted patient, we can guarantee the integrity of this part of the data.” Does this mean no vital signs were missing? If so, please state this in the methods. For laboratory tests, the authors say they used median imputation, which is reasonable, although multiple imputation would be preferable. However, of the scores tested, only MEDS/SOFA/APACHE-II use lab data. Again, I would state this in the methods for clarity if this was the rationale for imputation, otherwise no imputation would be required for development/validation of SSEPS.

Thanks for your suggestions and sorry for our unclear response. We have added statement as follow:

“*There is no missing data on vital signs and consciousness, which are basic information of each admitted patients, thus, no imputation would be required for development/validation of SSEPS.*” in ***Statistical analysis*** *section*.

• Calibration: Calibration on the development set is reported in the abstract and Figure 1. As per my previous comment, calibration on the validation set should be presented.

Thank you very much for your careful guidance, we have presented calibration on the validation set in our Figure 1B.

• Strength/limitations bullet points: Although mentioned in the discussion, the authors do not include the small sample size (236 patients in the validation set) here.

Thanks for your advice, we mentioned this limitation about small sample size in our *Strength/limitations bullet points*.

“The limitations of this study are small sample size (only 236 patients in the validation group) and lack of external validation dataset, which may limit the rigour of the study and confidence in the findings.