PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	DECAF score as a mortality predictor for acute exacerbation of chronic obstructive pulmonary disease: a systematic review and meta-analysis
AUTHORS	Huang, Qiangru; He, Chengying; Xiong, Huaiyu; Shuai, Tiankui; Zhang, Chuchu; Zhang, Meng; Wang, Yalei; Zhu, Lei; Lu, Jiaju; Jian, Liu

VERSION 1 - REVIEW

REVIEWER	Shouhao Zhou Penn State College of Medicine
REVIEW RETURNED	04-Mar-2020

GENERAL COMMENTS	This is a study of meta-analysis to assess the prognostic effect of DECAF score for patients with AECOPD. Requested by editor, my following review is with a particular emphasis on the statistical methods and analyses. Major concerns: 1. No reference was provided in statistical analysis. It was hard to tell how SROC approach was applied. If it was standard SROC approach, then it ignored heterogeneity between studies. This should be acknowledged as a major limitation of the study, or the authors should employ hierarchical summary receiver operating characteristic (HSROC) model (Ref: PMID: 11568945). 2. It was unclear how "Spearman's correlation coefficient was used to evaluate the threshold of the DECAF score prognostic accuracy." How was Spearman's correlation coefficient calculated in this case? 3. Optimal cutoff can be determined using meta-analysis but how the optimal cutoff was justified in the manuscript was incorrect. The authors can either remove the relevant statements or use correct statistical method. A suggestion is to use R package "diagmeta" (Ref: PMID: 27520527).

REVIEWER	Carlos Echevarria RVI, Newcastle I was an author on the DECAF validation study and the DECAF implementation study.
REVIEW RETURNED	13-Mar-2020

GENERAL COMMENTS	BMJ open DECAF meta-analysis Summary This study is in an interesting and important area. The results are mostly well considered and well presented. However, there are several areas that would need improving before publication. In particular, I think the authors need to be far more rigorous in their assessment of the risk of bias in the different studies. Introduction P4,L21. I would avoid the term “early warning scores” as these tend to be used for things like the “National Early Warning Score” (NEWS) which are used for inpatient monitoring of patients, rather than risk assessment at the point of presentation. I would suggest the term “prognostic scores”. P4,L21 suggest the term “high-risk” is not appropriate in the sentence “change strong indicator of identifying high-risk” as the authors are referring to risk stratification (high and low risk) and not just high risk patients. It would be worth discussing prognostic scores in terms of derivation, internal and external validation, and implementation studies as per the TRIPOD guidelines. I am not convinced that looking for a single optimal cut-off makes clinical sense. Risk scores categorise patients into different risk groups. The sens/ spec/ PPV and NPV will vary depending on which cut-off you are using, and the optimal cut off depends on whether you are trying to identify low risk or high risk patients. It would make sense to look at the cut offs for low risk and high risk groups separately, as well as the proportions identified by the cut-offs (for example, one of the benefits of the DECAF score is that it correctly identifies a large proportion of patients as low risk, which is of relevance if trying to reduce in-hospital bed stay). It should be noted that some of the other scores that are mentioned were not derived to look at COPD exacerbation, therefore it is unsurprising that they perform poorly- they may still be very useful in the context in which they were derived. P5 L7 I do not think the aim of this study is “to provide an effective and feasible prognostic tool…” and this needs amending. Methods The study could be much improved by expanding on the strengths and weakness and potential biases in each of the studies with the use of a reporting framework such as the TRIPOD checklist. Furthermore, I do not think that the current interpretation of the DECAF studies are correct in terms of their designs. For example there are some studies in which the data collection started before the publication of the DECAF score. It is not possible for these to be prospective trials. Furthermore, the eMRCD score is the strongest predictor and needs to be collected at the time either as part of the research or because it has been embedded into usual care. This may account for some of the variable performance of DECAF in different studies and should be addressed.

	Results Table 1 is useful but contains a typo- the eMRCD 5b score should be weight 2, and eMRCD 5a should be weighted 1. Table 2 is of use but the study design is not accurate for all of these studies as already mentioned. I would include a table in this section looking at the different studies and their biases for the DECAF scores studies in terms of the tripod statement, or another similar checklist to highlight risks and biases directly related to prognostic research looking at prediction tools. For example, were patients identified prospectively and consecutively, is the sample size what would be expected for the size of the hospital and recruitment time period, did COPD patients have confirmed COPD by a clinician and spirometry, what were the rates of missing data at the variable level and at the score level, how was missing data dealt with. etc… The results of the different scores in terms of the AUROC is helpful and of use, including the number of studies and the number of patients. I would consider looking at table 4 and the cut off in terms of high risk and low risk patients, and what is desirable. For example, for the clinician the NPV is of more importance in the low risk group and PPV is of more importance in the high risk group. Again table 5 does not make sense in as it is unclear what cut-off has been used- are we looking at high risk or low risk patients? Some scores may be good at identifying high risk patients, but may be poor at identifying low risk patients. Having separate results for inpatient and 30 day mortality is helpful and should remain. Figure 2- as mentioned before, it may make sense to have separate figures for high risk and low risk groups with different cut offs, or to more clearly label the rationale for the selected cut offs in this figure. Figure 3- I would have a more detailed legend explaining what this figure means. Supplementary figure- I think the risk of bias is integral and should appear in the main paper. However, I would suggest that the bias assessment is revisited. For example, in order to say that patient selection was of high quality, patients would have to have clearly defined COPD, be consecutively identified, etc, and I do not think this was the case for many of the studies. Similarly, for the index test, I am doubtful that many had the DECAF score collected correctly (in particular the eMRCD score) unless it was clearly prospectively planned. Again, those studies whose data collection predated the publication of DECAF can not have correctly collected the data. Instead of a separate table looking at biases based on the TRIPOD score, the biases could be included in this figure. P7, L34 The AUC cut-offs here are later contradicted. Here cut offs are described as moderate of they are 0.7-0.9 but later are described in the discussion as “very reliable” if above 0.8. For a clinical prediction tool I would retain the latter. Discussion P11 L34, modified DECAF and APACHE II are described as similar and in the next sentence DECAF is described as being better than these scores. This needs amending/ clarifying. I think the discussions needs to include information about the biases of the different studies, as per the TRIPOD statement.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name

Shouhao Zhou

Institution and Country
Penn State College of Medicine

Please state any competing interests or state ‘None declared’:
None Declared

Please leave your comments for the authors below
This is a study of meta-analysis to assess the prognostic effect of DECAF score for patients with AECOPD. Requested by editor, my following review is with a particular emphasis on the statistical methods and analyses.

Major concerns:
1. No reference was provided in statistical analysis. It was hard to tell how SROC approach was applied. If it was standard SROC approach, then it ignored heterogeneity between studies. This should be acknowledged as a major limitation of the study, or the authors should employ hierarchical summary receiver operating characteristic (HSROC) model (Ref: PMID: 11568945).
Response:
Thank you very much for your constructive comments and suggestions.
As suggested by reviewers, the hierarchical summary receiver operating characteristic (HSROC) model is indeed more suitable for analyzing our data. The HSROC model is appropriate and agile for diagnostic test, and could be used as an extension and improvement of the traditional SROC. We have changed the SROC method to the HSROC model, and revised the statements in the Method section of Data Synthesis and Analysis and the corresponding results (Figure 3). In addition, we also added references in this part.
[Method section of Data Synthesis and Analysis; second paragraph]: The mixed bivariate random-effects regression model was used to analyze and pool the diagnostic accuracy measurements across studies18. To derive summary estimates, we plotted estimates of the observed sensitivities and specificities for each test in forest plots and hierarchical summary receiver operating characteristic (HSROC) curves derived from individual study results19, 20. These results were plotted using HSROC curves with 95% confidence and prediction regions. Additionally, pooled sensitivity (SEN), specificity (SPE), diagnostic odds ratio (DOR), positive likelihood ratio (PLR), and negative likelihood ratio (NLR) were calculated21. The AUC was also calculated to show the prognostic performance of DECAF. In clinical practice, tests with AUC above 0.8 are considered to be very reliable22.

2. It was unclear how "Spearman’s correlation coefficient was used to evaluate the threshold of the DECAF score prognostic accuracy." How was Spearman’s correlation coefficient calculated in this case?
Response:
Thank you for your kind comments.
In diagnostic studies, heterogeneity might be caused by threshold effect or non-threshold effect due to the different cutoff values used in different studies. Therefore, we evaluated the heterogeneity caused by threshold effect by calculating the Spearman correlation coefficient between the sensitivity and

false-positive rate. If the Spearman correlation coefficient was greater than or equal to 0.6 ($p<0.05$), there was a threshold effect. Cochran's Q test was used to evaluate the heterogeneity caused by non-threshold effect. Deeks' funnel plot was applied to analyze publication bias. The meta-analysis results were presented by forest plots.

As reviewer's suggested, to clarify the statements, we revised the statement as follows: To assess the heterogeneity from the threshold effect, the Spearman correlation coefficient between the logit of sensitivity and the logit of (1− specificity) was computed to assess the threshold effect on the prognostic accuracy of DECAF score. If the Spearman correlation coefficient was greater than or equal to 0.6 ($p<0.05$), there was a threshold effect. The Deek's funnel plot asymmetry test was used to assess for publication bias, when the included studies were greater than 10 studies. [the Method section of Data Synthesis and Analysis; fourth paragraph]

3. Optimal cutoff can be determined using meta-analysis but how the optimal cutoff was justified in the manuscript was incorrect. The authors can either remove the relevant statements or use correct statistical method. A suggestion is to use R package "diagmeta" (Ref: PMID: 27520527).
Response:
Thank you for your kind comments and suggestions.
We removed the relevant statements about looking for the single optimal cut-off value of DECAF score. We realized that in clinical practice, it is more important to a separately ssess the cut-off values for risk stratification in low-risk and high-risk groups. Thus, we adjusted the statements as "explored the effectiveness of different cutoff values in risk stratification of AECOPD patients". And we also revised corresponding contents in the Introduction, Results, Discussion, and Conclusion Section.

Reviewer: 2
Reviewer Name

Carlos Echevarria

Institution and Country

RVI, Newcastle

Please state any competing interests or state 'None declared':
I was an author on the DECAF validation study and the DECAF implementation study.

Please leave your comments for the authors below
BMJ open DECAF meta-analysis
Summary
This study is in an interesting and important area. The results are mostly well considered and well presented. However, there are several areas that would need improving before publication. In particular, I think the authors need to be far more rigorous in their assessment of the risk of bias in the different studies.

Introduction
1. P4, L21. I would avoid the term "early warning scores" as these tend to be used for things like the "National Early Warning Score" (NEWS) which are used for inpatient monitoring of patients, rather than risk assessment at the point of presentation. I would suggest the term "prognostic scores".
Response:
Thanks for the reviewer's kind suggestion.

We carefully revised the statement according to the reviewers' comments, and adjusted the term “early warning scores” to “prognostic scores” throughout the text.

2. P4, L21 suggest the term “high-risk” is not appropriate in the sentence “change strong indicator of identifying high-risk” as the authors are referring to risk stratification (high and low risk) and not just high risk patients.
Response:
Thanks for your kind suggestion.
We revised the inappropriate statement according to the reviewers' suggestion, and we adjusted the statement to “Prognostic scores can provide a strong indicator for risk stratification and assist in clinical management… (P4, L9)”. All of these inappropriate statements have been corrected in the revised manuscript.

3. It would be worth discussing prognostic scores in terms of derivation, internal and external validation, and implementation studies as per the TRIPOD guidelines.
Response:
The TRIPOD guideline is a scientific guideline for diagnostic or prognostic original studies. As a secondary study, all methods of this systematic review and meta-analysis followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. On the basis of compliance with PRISMA, some modifications have been made with reference to TRIPOD guidelines. In the Introduction Section, we originally descripted the background of DECAF in three aspects, including the definition, component of test index, and the clinical application. Based on the reviewer’s suggestion, we found that our original statements were insufficient to describe the background of the DECAF score, so we added the following statements for explanation: “The DECAF score showed promising performance in derivative studies, and was superior to other prognostic tools for AECOPD patients6. The UK National COPD audit recommends that DECAF scores be recorded for AECOPD patients. However, it is also pointed out that the application of DECAF score still needs evidence and validation8. In addition, the prognosis value of DECAF score is still unclear and needs to be verified, which is essential to prove the generalization of prognosis scores.” [The second paragraph of the Introduction Section]

4. I am not convinced that looking for a single optimal cut-off makes clinical sense. Risk scores categorise patients into different risk groups. The sens/ spec/ PPV and NPV will vary depending on which cut-off you are using, and the optimal cut off depends on whether you are trying to identify low risk or high risk patients. It would make sense to look at the cut offs for low risk and high risk groups separately, as well as the proportions identified by the cut-offs (for example, one of the benefits of the DECAF score is that it correctly identifies a large proportion of patients as low risk, which is of relevance if trying to reduce in-hospital bed stay).
Response:
Thank you very much for your detailed explanation and suggestions.
Through the reviewer's explanation, we do realize that the statement of optimal cut-off value is inappropriate and lacks clinical value. In the revised manuscript, we deleted all the related statements about looking for a single optimal cut-off value. And the subgroup analysis conducted based on different cut-off values of DECAF score is aimed to explore the effectiveness of different cut-off values in risk stratification of AECOPD patients, and to evaluate the specific prognostic effect of each cut-off value. Thus, we adjusted the statements as “This systematic review and meta-analysis evaluated the association between DECAF scores and the prognosis of AECOPD patients, assessed the specific predictive and prognostic value of DECAF scores, and explored the effectiveness of different cut-off values in risk stratification of AECOPD patients.” [The third paragraph of Introduction Section].

5. It should be noted that some of the other scores that are mentioned were not derived to look at COPD exacerbation, therefore it is unsurprising that they perform poorly- they may still be very useful in the context in which they were derived.
Response:
In the derivation study, DECAF showed strong performance and was superior to other tools for patients with AECOPD. However, the UK National COPD audit also pointed out that it still required validation and evidence. To further assess and validate the clinical value of DECAF scores, we compared the prognostic value of DECAF to other commonly used prognostic scores, including the modified DECAF, CAPS, CURB-65, and APACHE II scoring systems.
As reviewer's suggestion, we added the following sentences to clarify the statement.
"Although these scores are not designed or proposed for AECOPD, they are still commonly used in clinical practice for the prediction and prognostic evaluation of AECOPD patients". [The third paragraph of Introduction Section]

6. P5 L7 I do not think the aim of this study is "to provide an effective and feasible prognostic tool…" and this needs amending.
Response:
This systematic review and meta-analysis evaluated the association between DECAF scores and the prognosis of AECOPD patients, assessed the specific predictive and prognostic value of DECAF scores, and explored the effectiveness of different cutoff values in risk stratification of AECOPD patients. To further assess the clinical value of DECAF scores, we compared the test to other commonly used prognostic scores.
As reviewer's comment, we adjusted the statement of study purpose as "This study aimed to evaluate and validate the effectiveness of the DECAF score and improve the clinical course and outcome of AECOPD patients". [The third paragraph of Introduction Section]
We also adjusted the statement in Abstract as follows: This study was conducted to assess the association between DECAF scores (The Dyspnea, Eosinopenia, Consolidation, Acidemia, and Atrial Fibrillation) and the prognosis of patients with acute exacerbation of chronic obstructive pulmonary disease (AECOPD), and to evaluate the specific predictive and prognostic value of DECAF scores, and to explore the effectiveness of different cut-off values in risk stratification of AECOPD patients.

Methods
7. The study could be much improved by expanding on the strengths and weakness and potential biases in each of the studies with the use of a reporting framework such as the TRIPOD checklist.
Response:
Thank you for your kind comments.
As a secondary study, all methods of this systematic review and meta-analysis followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. As the original studies included in this meta-analysis are diagnostic or prognostic studies, according to the Cochrane handbooks, the quality assessment and risk of bias in the included studies should be assessed by the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2). The TRIPOD statement is a scientific guideline for diagnostic or prognostic original studies, hence, we mainly followed PRISMA guidelines and QUADAS-2 for this meta-analysis.
Considering the reviewer' suggestion, we expanded the quality assessment in the Methods Section. We also added a new table (Table 3: The Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) for included studies) to evaluate the potential biases in each of the studies. And the corresponding Results Section and Discussion Section were also revised.
In the Quality Assessment part of Methods Section, we adjusted the following statements to clarify and expand the quality assessment and risk of bias. "Two review authors (J Liu and J Lu) independently applied the guidelines of the PRISMA statement16 to evaluate each involved study. The Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) was conducted by two independent authors (J Liu and J Lu) to assess the quality and risk of bias for diagnostic or prognostic

studies17. In case of any inconsistency, all authors reach an agreement through discussion. The quality and risk of bias were assessed from two perspectives, including bias risk and applicability concerns, and evaluated from four aspects, including patient selection, index test, reference standard, and flow and timing."

8. Furthermore, I do not think that the current interpretation of the DECAF studies are correct in terms of their designs. For example there are some studies in which the data collection started before the publication of the DECAF score. It is not possible for these to be prospective trials. Furthermore, the eMRCD score is the strongest predictor and needs to be collected at the time either as part of the research or because it has been embedded into usual care. This may account for some of the variable performance of DECAF in different studies and should be addressed.
Response:
Thank you for your detailed explanation and suggestions.
According to the comments of the reviewers, we realized the importance of DECAF score collection. We carefully examined and reviewed the full text of all included studies to check the study's design. The following are examples of five studies which we made adjustments in the item of Study Design in Table 2 (Characteristics of included studies).
•The study conducted by Shafuddin 2018 was a retrospective design, they used data from two prospective cohorts of patients hospitalised for a primary diagnosis of an exacerbation of COPD, and they collected the DECAF score by compiled admission data.
•The study conducted by Xu 2017 was a case-control design, they derived patients into survival and non-survival groups, however, they reported the collection of DECAF was conducted within 24 hours after admission for all of patients.
•In the study conducted by Echevarria 2016, they reported that "In the internal validation cohort hospitals, the DECAF indices are recorded as part of routine practice. This allowed the period of the study to be extended retrospectively to enhance recruitment; patients were primarily identified from a broad coding records search (discharge codes). In the external validation cohort to identify consecutive admissions of patients with AECOPD, all medical admissions were screened prospectively", their study has a retrospective and prospective design, and the collection of DECAF was recorded as part of routine practice.
•The study by Yousif 2016 was also a retrospective and prospective design, they reported that either retrospectively (174 patients) by reviewing the hospital's records or prospectively (90 patients), and the collection of DECAF was compiled by admission data.
•The study by Steer 2012 was a prospective design, they recruited patients without acknowledgements of the main outcomes (e.g. mortality of in-hospital). As far as we known, the concept of DECAF score was first proposed in this very study. They reported that "Socio-demographic and clinical data were collected on admission." and "Independent predictors of outcome were identified by logistic regression analysis and incorporated into a clinical prediction tool". Thus, the collection of DECAF in this study was analyzed and compiled by clinical data.
We added the statements for collection of DECAF into data extraction and added this item into Table 2 (Characteristics of included studies). "With regard to the collection of DECAF score, eight studies collected the score on admission9, 27, 30, 32-34, 38, 40, one reported that the collection was pre-specified in the original study protocol26, one was collected within 24 hours after admission 35, one recorded DECAF score as part of routine practice28, and the other six reported that the DECAF score was compiled based on admission data6, 29, 31, 36, 37, 39." [The Study Characteristics Section of Results Section].
We also added corresponding statements in the heterogeneity analysis part of Discussion Section. "The biases between included studies can also lead to heterogeneity. The DECAF score needs to be collected at admission or pre-specified in the original study protocol. However, the collection of DECAF score varied between the included studies, which may result in variable performance of DECAF." [The penultimate paragraph of the Discussion]

Results
9. Table 1 is useful but contains a typo- the eMRCD 5b score should be weight 2, and eMRCD 5a should be weighted 1.
Response:
Thanks for reviewer's kind mention.
We have revised the typo in Table 1, and we also rechecked all of the Tables and Figures to avoid typo errors.

10. Table 2 is of use but the study design is not accurate for all of these studies as already mentioned.
Response:
Table 2 has been adjusted to make the study design more accurate. Please refer to the response to question 8 for details.

11. I would include a table in this section looking at the different studies and their biases for the DECAF scores studies in terms of the tripod statement, or another similar checklist to highlight risks and biases directly related to prognostic research looking at prediction tools. For example, were patients identified prospectively and consecutively, is the sample size what would be expected for the size of the hospital and recruitment time period, did COPD patients have confirmed COPD by a clinician and spirometry, what were the rates of missing data at the variable level and at the score level, how was missing data dealt with. etc…
Response:
Thanks for reviewer's kind suggestion and detailed explanation.
By reviewer's suggestion, we added a new table in Results Section (Table 3: The Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) for included studies) to assess the quality and bias of each included studies. The assessed items are based on the QUADAS-2 guidelines.
We also added statements about the risk of bias in the Methodological Quality and Risk of Bias of Results Section.
Only one study was a case-control design without blinding statements, which could not prevent the occurrence of observer bias, thus the risk of bias was related high 35. All studies included patients diagnosed with AECOPD, and eight studies reported consecutive enrollment 6, 9, 26-28, 30, 34, 40. Most of studies included did not pre-specify the cut-off value for risk stratification. Since the main outcome is the mortality of AECOPD patients, for which the reference standard is survival or non-survival, all included studies met the low-risk criteria of the reference standard items. However, the included studies yielded different baseline characteristics in the included population, which affected patient selection, flow, and timing. The quality and bias of each included studies was shown in Table 3, and the summary figures of risk of bias were shown in Figs. S1 and S2.

12. The results of the different scores in terms of the AUROC is helpful and of use, including the number of studies and the number of patients.
Response:
According to reviewer's suggestion, we added the number of studies and the number of patients in the Results Section of the subgroup analysis based on different cut-off values (Results Section of Subgroup Analysis; the second paragraph). And in corresponding Table 5 (Subgroup analysis of the prognostic value of DECAF based on different variables) we also included the number of studies and patients.

13. I would consider looking at table 4 and the cut off in terms of high risk and low risk patients, and what is desirable. For example, for the clinician the NPV is of more importance in the low risk group and PPV is of more importance in the high risk group.
Response:
Thanks for the reviewer's constructive comments and suggestions.

As suggested by the reviewer, in clinical practice, the greater the positive likelihood ratio (PLR), the greater the probability of a true positive when the test result is positive. The smaller the negative likelihood ratio (NLR), the more likely it is to be true negative when the test result is negative. In clinical practice, PLR is of more importance in the stratification of high-risk group and the NLR is of more importance in the low-risk patients.
We added the descriptions of PLR and NLR for the results of deferent cut-off values in the second paragraph of Subgroup Analysis of Results Section.
We also added a paragraph in Discussion Section as follows (the sixth paragraph of Discussion Section): “In clinical practice, the greater the PLR value, the greater the likelihood of true positive when the test result is positive; the smaller the NLR value, the greater the likelihood of true negative when the test result is negative. PLR is more important in stratification of high-risk groups, while NLR is more important in low-risk groups. From the results, the NLR was very small, 0.31, which indicated that the DECAF score could correctly identify most AECOPD patients as a low-risk group. For the cut-off value from 2 to 4, the PLR value increased from 1.80 to 3.80, indicating that with the increase of the cut-off value, the risk stratification of the DECAF score in high-risk groups increased significantly.”

14. Again table 5 does not make sense in as it is unclear what cut-off has been used- are we looking at high risk or low risk patients? Some scores may be good at identifying high risk patients, but may be poor at identifying low risk patients.
Response:
Indeed, as recommended by the reviewers, it is more meaningful to evaluate different cut-off value for high risk or low risk patients, but the current number of original studies is insufficient, and different original studies use different cutoff values for these scores. Taking modified DECAF score as an example, only 3 articles reported relevant results, and the cutoff values were different. Therefore, as a meta-analysis, it is difficult for us to perform subgroup analysis based on different cutoff values for these scores. Instead, we choose to compare the comprehensive results of each score and obtain the preferred score by comparison.
We supplement this part of the description in the article's limitations section, hoping to have more original research to support further research, and explore the role and effect of different scores in the identification of high risk or low risk patients. “Thirdly, because of the lack of original research comparing DECAF with other predictive scores, we can only compare the predictive value of DECAF and other predictive scores to AECOPD patients in general. With the increase of related original research, it is possible to further explore the effectiveness of different prognostic scores in risk stratification of AECOPD patients.”[The last paragraph of the Discussion]

15. Having separate results for inpatient and 30 day mortality is helpful and should remain.
Response:
As the reviewer suggested, we kept the separate results for in-hospital and 30-day mortality.

16. Figure 2- as mentioned before, it may make sense to have separate figures for high risk and low risk groups with different cut offs, or to more clearly label the rationale for the selected cut offs in this figure.
Response:
Only a few of the included studies clearly defined the risk level of the population corresponding to the cut-off value, while most studies did not pre-specify the cut-off value for risk stratification. Therefore, Figure 2 showed the comprehensive sensitivity and specificity of DECAF for the prediction of mortality in AECOPD, without distinguishing between high risk and low risk groups.
The cut-off values are extracted from the included studies. Most of the articles reported the optimal cut-off value of DECAF and its corresponding indicators such as sensitivity and specificity. There are also some articles that show the sensitivity and specificity under different cut off values. For such articles, we choose the optimal cut-off value. For studies that did not report cut-off values, we extracted relevant effect quantities (e.g. tp, fp, tn, fn) and included them in the calculation of pooled

sensitivity, specificity, and AUC, etc. The cut-off value of each included study is shown in the "DECAF Cut-off value" item in Table 2.

17. Figure 3- I would have a more detailed legend explaining what this figure means.
Response:
Thanks for your suggestion, we added a detailed legend to explain what Figure 3 means.
The HSROC curves was conducted which plots sensitivity versus specificity. All studies were presented as a circle and plotted with the HSROC curve. The summary point (red box) indicates that the summary sensitivity was 0.76 and the summary specificity was 0.76. The summary results are displayed as the 95% confidence region and 95% prediction region in the HSROC curve plot. The size of the marker is scaled according to the total number of patients in each study.

18. Supplementary figure- I think the risk of bias is integral and should appear in the main paper. However, I would suggest that the bias assessment is revisited. For example, in order to say that patient selection was of high quality, patients would have to have clearly defined COPD, be consecutively identified, etc, and I do not think this was the case for many of the studies. Similarly, for the index test, I am doubtful that many had the DECAF score collected correctly (in particular the eMRCD score) unless it was clearly prospectively planned. Again, those studies whose data collection predated the publication of DECAF can not have correctly collected the data. Instead of a separate table looking at biases based on the TRIPOD score, the biases could be included in this figure.
Response:
Thanks for reviewer's kind suggestion.
As suggested by the reviewer, we re-checked and revised quality and bias assessment for each included study. And we also added the followed statements in the heterogeneity assessment part of Discussion Section (the penultimate paragraph of Discussion Section). "The biases between included studies can also lead to heterogeneity. The DECAF score needs to be collected at admission or pre-specified in the original study protocol. However, the collection of DECAF score varied between the included studies, which may result in variable performance of DECAF. In addition, different included studies yielded different baseline characteristics in the included population, which affected patient selection and also led to the different selection of cut-off value between studies."
We added a new table to assess the quality and bias of each included studies (Table 3: The Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) for included studies). According to the Cochrane handbooks, for diagnostic or prognostic meta-analysis, the QUADAS-2 guideline is recommended for the assessment of quality and bias of included studies. The assessed items in Table 3 are based on the QUADAS-2 guidelines. The content of Table 3 is more specific than the content of Figures S1 and S2, and due to the limitation of the length of the article, we chose to put Table 3 in the main text and leave Figures S1 and S2 in the supplementary file for readers.

19. P7, L34 The AUC cut-offs here are later contradicted. Here cut offs are described as moderate of they are 0.7-0.9 but later are described in the discussion as "very reliable" if above 0.8. For a clinical prediction tool I would retain the latter.
Response:
We have unified the statements about the AUC cut-offs as "In clinical practice, test with AUC greater than 0.8 is considered to be very reliable".

Discussion
20. P11 L34, modified DECAF and APACHE II are described as similar and in the next sentence DECAF is described as being better than these scores. This needs amending/ clarifying.
Response:
Thanks for reviewer's kind mention.

We clarified the related statements as “Quantitative analysis demonstrated that elevated DECAF scores were significantly associated with high mortality risk. In other potential scoring systems, compared with the survivor group, the results showed that only the modified DECAF and APACHE Ⅱ scores increased in the non-survivor group In the accuracy analysis, DECAF scores showed a reliable prognostic accuracy for both in-hospital and 30-day mortality. When the prognostic value was compared with other prognostic scores, DECAF scores showed better prognostic accuracy and stable clinical value in predicting the in-hospital mortality and 30-day mortality of patients with AECOPD.”[The second paragraph of Discussion Section]

21. I think the discussions needs to include information about the biases of the different studies, as per the TRIPOD statement.
Response:
Thanks for reviewer’s kind suggestion.
We conducted a subgroup analysis to assess sources of heterogeneity and bias, and analyzed heterogeneity in the penultimate paragraph of section Discussion. According to the suggestions of Reviewer, we supplemented the discussion on biases of the different studies in this part. For details, see the yellow mark of the penultimate paragraph in the section Discussion.

VERSION 2 – REVIEW

REVIEWER	Shouhao Zhou Penn State College of Medicine
REVIEW RETURNED	14-May-2020

GENERAL COMMENTS	All my concerns have been addressed. PS: It was inconsistent that both "cutoff" and "cut-off" were used in the manuscript.

REVIEWER	C Echevarria RVI hospital, Newcastle, NE1 4LP, United Kingdom I was an author on some of the work cited in the study.
REVIEW RETURNED	18-May-2020

GENERAL COMMENTS	The authors have addressed almost all of the comments I have made. Some of the additional elements including table 3 have further improved the work. I am satisfied with the changes that have been made, and only have a few outstanding comments that require attention. Page 5, line 44. The Stone reference (8) pre-dates subsequent work. The DECAF validation publication followed this reference, and therefore the DECAF score has been validated by this and by the other publications cited. Of note, the Hospital at Home RCT, was an implementation study, and this has not been mentioned or cited. In prognostic research it is important, though unusual, for a prognostic score to be assessed in clinical practice. The introduction ought to mention that the DECAF score has undergone derivation, internal and external validation, and implementation. The third study (DECAF implementation) only included low risk patients for Hospital at Home, but the mortality rate this group was low confirming that

	DECAF can be used in practice. This would be worth mentioning the introduction, as well as the discussion. DECAF implementation RCT (Echevarria C, Gray J, Hartley T, et al. Home treatment of COPD exacerbation selected by DECAF score: a non-inferiority, randomised controlled trial and economic evaluation. Thorax 2018.) I think table 3 is extremely helpful addition to the paper. However, I would be a little cautious answering yes to question 1 for some of these papers. In order to establish if the papers truly had consecutive recruitment, I would suggest looking at the number of patients who were recruited per hospital over the time period. Is the number what you would expect for the time period and the size of the hospital? Furthermore, if they obtained written consent (which is not necessarily for observational studies) this inadvertently excludes the most well and the most unwell patients and introduces bias. For question 4 I would look at whether patients all had spirometry confirmed COPD (in other words, did the researchers have the requirement for obstructive spirometry at some point in the passed within their inclusion criteria). We know that many patients labelled with COPD do not actually have COPD, and therefore to identify this population spirometry is absolutely necessary. Finally, if the authors used ICD codes alone (or other coding systems) to identify patients, then I this should seen as introducing bias. ICD codes may miss patients with COPD who have co-existent pneumonia, and also include patients who do not have COPD exacerbation. Therefore, if authors have used coding to identify patients then the notes should have been reviewed by respiratory specialists to confirm the diagnosis of an exacerbation of COPD. I see at least one paper (Steer) that perhaps should be Yes for question 5: “If the index test is always conducted and interpreted before the reference standard, this item can be rated “yes.”” Echevarria 2016: “Although retrospective collection of data may bias results, this risk was mitigated as the DECAF indices were collected as part of routine clinical practice in the participating hospitals, the researchers extracting data were blinded to outcome and case ascertainment and outcome were similar to the prospective external cohort.” There is also further details about blinding in the online supplement. I would suggest this addresses question 7 and 9 which are currently marked as “unclear”.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name

Shouhao Zhou

Institution and Country

Penn State College of Medicine

Please state any competing interests or state ‘None declared’:
None declared

Please leave your comments for the authors below
All my concerns have been addressed.

PS: It was inconsistent that both "cutoff" and "cut-off" were used in the manuscript.
Response:
Thanks for reviewer’s kind mention, we have unified the statement as “cutoff” in the revised manuscript.

Reviewer: 2
Reviewer Name

Carlos Echevarria

Institution and Country

RVI hospital, Newcastle, NE1 4LP, United Kingdom

Please state any competing interests or state ‘None declared’:
I was an author on some of the work cited in the study.

Please leave your comments for the authors below
The authors have addressed almost all of the comments I have made. Some of the additional elements including table 3 have further improved the work. I am satisfied with the changes that have been made, and only have a few outstanding comments that require attention.
Response:
Special thanks to you for your valuable comments and suggestions. We appreciate for your warm and kind work earnestly, and hope that the correction will meet with approval.

1. Page 5, line 44. The Stone reference (8) pre-dates subsequent work. The DECAF validation publication followed this reference, and therefore the DECAF score has been validated by this and by the other publications cited. Of note, the Hospital at Home RCT, was an implementation study, and this has not been mentioned or cited. In prognostic research it is important, though unusual, for a prognostic score to be assessed in clinical practice. The introduction ought to mention that the DECAF score has undergone derivation, internal and external validation, and implementation. The third study (DECAF implementation) only included low risk patients for Hospital at Home, but the mortality rate this group was low confirming that DECAF can be used in practice. This would be worth mentioning the introduction, as well as the discussion.
Response:
Thanks for reviewer’s kind suggestion and detailed explanation.
As reviewer’s suggested, to clarify the statements, we revised the statement as follows: The Dyspnea, Eosinopenia, Consolidation, Acidemia, and Atrial Fibrillation (DECAF) score is a risk stratification tool designed to predict risk of death in AECOPD patients6, and can be easily applied at the bedside to guide treatment, such as hospital at home for low-risk patients7. [the second paragraph of Introduction Section]. In 2014, the UK National COPD audit recommends that DECAF scores be recorded for AECOPD patients9. Subsequently, an increasing number of original studies conducted derivation, internal and external validation, and implementation of the DECAF score. The

prognostic value of DECAF score still needs to be further verified by the methods of systematic review and meta-analysis, which is essential to prove the generalization of prognosis scores. [the second paragraph of Introduction Section].
In addition, we also added relevant statements in Discussion Section: In a randomized controlled trial and economic evaluation study of DECAF implementation, the low-risk patients (DECAF 0 or 1) selected by DECAF were more cost-effective than the usual care, mainly manifested in a 5-fold reduction in the median of 90 days of hospitalization7. The study showed that the DECAF score was easily applied at the bedside to guide treatment, and about twice as many patients were eligible compared with earlier models7. It was safe, clinically effective, cost-effective to use DECAF score at home in low-risk patients, and preferred by most patients7. [the fifth paragraph of Discussion Section]

2. DECAF implementation RCT (Echevarria C, Gray J, Hartley T, et al. Home treatment of COPD exacerbation selected by DECAF score: a non-inferiority, randomised controlled trial and economic evaluation. Thorax 2018.)
Response:
Thanks for the help of reviewers, this paper enables us to have a better understanding of the application of DECAF in hospital at home. We quoted this paper in the Introduction and Discussion part of the revised manuscript, corresponding to reference 7.

3. I think table 3 is extremely helpful addition to the paper. However, I would be a little cautious answering yes to question 1 for some of these papers. In order to establish if the papers truly had consecutive recruitment, I would suggest looking at the number of patients who were recruited per hospital over the time period. Is the number what you would expect for the time period and the size of the hospital? Furthermore, if they obtained written consent (which is not necessarily for observational studies) this inadvertently excludes the most well and the most unwell patients and introduces bias.
Response:
Thank you for your detailed explanation and suggestions.
Among the seven studies rated yes in question 1, three are multicenter observational studies (Echevarria 2019, Echevarria 2016, and Steer 2012) and four are single center studies (Shi 2019, Sangwan 2017, Zidan 2015, and Nafae 2014). As for the multicenter observational studies, in Echevarria 2016, they reported that “Based on an expected sensitivity of 70%, an SE of the estimate of sensitivity of 5% required a minimum of 840 patients in both the internal and external validation cohorts.”, and the number of recruited patients met expectations. In Steer 2012, they reported that all of the recruited patients were consecutive, and they also demonstrated that “Approval was granted by the local National Health Service Research Ethics Committee who advised that individual patient consent was not required”. In Echevarria 2019, they reported that “The DECAF derivation and validation cohorts are composed of 2645 consecutive admissions of unique patients with ECOPD to six UK hospitals with preadmission obstructive spirometry (ISRCTN 13946813 and 29082260)”. Thus, we think these three studies are appropriate for question 1 to mark “yes”. As for other four single center studies, they only recruited patients in one hospital, and they also specifically reported that they recruited a consecutive sample of patients.
There were only two studies reported the statements about ethical approval. Steer 2012: Approval was granted by the local National Health Service Research Ethics Committee who advised that individual patient consent was not required. Echevarria 2016: Ethical approval was granted by the local research ethics committee. And other five studies did not report any statements about written consent or ethical approval.
Although the written consent of observational studies may introduce bias in the selection of patients, whether a patient signs written consent is beyond the control of the observer. For mild, moderate, and severe patients, we believe that there is no statistical difference, and it can be considered a random event, so the effect of this bias on the results can be reduced or ignored. In addition, in order to

distinguish other studies that did not explicitly indicate a consecutive or random sample of patients enrolled, those studies report a consecutive sample of patients should be rated as “yes” in question 1.

4. For question 4 I would look at whether patients all had spirometry confirmed COPD (in other words, did the researchers have the requirement for obstructive spirometry at some point in the passed within their inclusion criteria). We know that many patients labelled with COPD do not actually have COPD, and therefore to identify this population spirometry is absolutely necessary. Finally, if the authors used ICD codes alone (or other coding systems) to identify patients, then I this should seen as introducing bias. ICD codes may miss patients with COPD who have co-existent pneumonia, and also include patients who do not have COPD exacerbation. Therefore, if authors have used coding to identify patients then the notes should have been reviewed by respiratory specialists to confirm the diagnosis of an exacerbation of COPD.
Response:
Thank you for your detailed explanation and suggestions.
Through the reviewer's explanation, we do agree that spirometry confirmed COPD is necessary, and used ICD codes alone may also introduce bias in the selection of patients. For question 4, we re-checked all of the included studies, most of studies which rated low risk in the question 4 reported the spirometry confirmed COPD. Diagnosis of AECOPD according to Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria supported by spirometry evidence of airflow obstruction (forced expiratory volume in one second (FEV1)/forced vital capacity (FVC) < 0.70) when clinically stable; with clinical criteria of exacerbation including increased dyspnea, increased sputum volume or sputum purulence. Three studies are currently rated as “Low risk” in question 4 (Bastidas 2018, Sangwan 2017, and Zidan 2015), although these studies demonstrated that they recruited patients with AECOPD, thsy did not report the spirometry confirmed statements, and the diagnosis criteria of AECOPD was unclear. Thus, after consulted with all of the authors in our study, we have changed it as “Unclear risk”. The changes were marked as yellow in Table 3. All changed items in Table 3 have been modified in Figures S1 and S2 of the supplementary.

5. I see at least one paper (Steer) that perhaps should be Yes for question 5: “If the index test is always conducted and interpreted before the reference standard, this item can be rated “yes.””
Response:
Thank you for your kind comment and suggestion.
As reviewer’s suggestion, we have re-checked the question 5 in each included study, we do realize that this question in the study by Steer should be yes. And we found that in the study by Shi 2019, the question 5 is also should answer yes, because they reported the collection time of DECAF was at admission, and they also pointed out the outcome measured time and the follow-up period. We have revised the results of question 5 in Table 3 and highlight these changes in yellow. This changes in Table 3 have also been modified in Figures S1 and S2 of the supplementary.

6. Echevarria 2016: “Although retrospective collection of data may bias results, this risk was mitigated as the DECAF indices were collected as part of routine clinical practice in the participating hospitals, the researchers extracting data were blinded to outcome and case ascertainment and outcome were similar to the prospective external cohort.” There is also further details about blinding in the online supplement. I would suggest this addresses question 7 and 9 which are currently marked as “unclear”.
Response:
Thanks for reviewer’s kind mention, we have carefully re-checked the relevant statements in Echevarria 2016. Through reviewer’s detailed explanation, we do agree that question 7 and 9 should be grated as low risk in the study conducted by Echevarria 2016. After consulted with all of the authors in this meta-analysis, we have revised the results in Table 3 and highlight these changes in yellow. We also re-checked the question 7 and 9 in other included studies, and no changes have

been made. All changed items in Table 3 have been modified in Figures S1 and S2 of the supplementary.

VERSION 3 – REVIEW

REVIEWER	C Echevarria RVI hospital, Newcastle, United Kingdom As previously stated, I was an author on the DECAF validation study.
REVIEW RETURNED	07-Jul-2020

GENERAL COMMENTS	I am satisfied with all the of the changes that have been made, and that the work is ready for publication.