PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Development of the Combined Assessment of Risk Encountered in Surgery (CARES) surgical risk calculator for prediction of post-surgical mortality and need for Intensive Care Unit admission risk – A Single-Centre Retrospective Study
AUTHORS	Chan, Diana; Sim, Eileen; Chan, Yiong Huak; Poopalalingam, Ruban; abdullah, hairil

VERSION 1 – REVIEW

REVIEWER	Iain Moppett University of Nottingham, UK
REVIEW RETURNED	09-Oct-2017

GENERAL COMMENTS	The authors present the development and local validation of a novel risk prediction tool for most types of adult surgery. They found high discriminatory ability and adequate calibration. Of note, the study is deliberately investigating the largely under-researched South East Asian population. The authors describe a sensible rationale for why the study was done (different population) and their inclusion of RDW as a biomarker. I would like to see a STROBE checklist or similar please. This may address some of my comments. The details of the modelling need some expansion, even if in a supplementary appendix. What were the univariate relationships (with n/N, OR, p) for each of the included variables. How were items selected or excluded? What were the selection criteria for the optimum model - AIC / BIC are probably appropriate here. Did the authors looks for interaction between terms? I am not sure what the authors mean by a rank model. I think what they have done is (in a fairly non-linear fashion) rounded the OR to create simpler co-efficients. On the one hand I see the benefits of the rounding approach (it makes explanation to patient and clinicians easier) but it isn't needed with current online calculators (and a calculator is needed due to the logistic equation). Also, why have they chosen to decrease the impact of the high OR items? The equation for predicting mortality from the score needs to be published otherwise it is impossible to replicate (or use) the results.

	I am confused by the H-L plots. The apparent goodness of fit seems to be strongly influenced by the high risk decile. It would be helpful to tabulate these data. I am assuming that the plot circle areas are proportional to group size. (I think there may be a typo in the Chi-sq probabilities. Both H-L graphs give same p, but different Chi-sq. My calculator gives 3a as 0.77) Looking at the figures, the calibration appears to be not great at low risk. This probably doesn't matter at individual patient level (I'm not sure there is a meaningful difference between very low risk groups) but does matter if used for benchmarking. Table 3 - I'm not sure this adds much. 88% of patients are going to score 3 by virtue of age alone. Is there a reason why the authors didn't compare their score with SORT, POSSUM etc.? The premise of the paper is that these are unvalidated in the SE Asian population. This would be an opportunity to address that. This is important as if they don't work this needs to be shown and known about. On minor note, how were the CI for AUROC calculated (there are several ways to do this). Iain Moppett Nottingham, UK

REVIEWER	Wei Zhu Stony Brook University, USA
REVIEW RETURNED	23-Oct-2017

GENERAL COMMENTS	This paper arises from the need of medical services and is of good value to medicine. Thee method described is also suffice, However there is one major short-coming of the study design, that is, the robustness of the training and testing sets should be examined. This can be easily done by randomly choose the training and testing sets a few times, and for each pair, repeat the procedures to see whether the variables selected are stable, and whether the classification performance are comparable. Such resampling results should be summarized for its distribution in a separate section.

REVIEWER	Nirav Shah St George's Hospital,United Kingdom
REVIEW RETURNED	26-Oct-2017

GENERAL COMMENTS	There are a number of gramattical and vocabulary based errors that need rectifying prior to publication. The reasons for the study do not appear to be clear. It is true that there is little/ no validation work for the currently used scoring systems in Asian populations, but the second reason that is dwelt upon is that these systems are unwieldy and require pre-op blood tests etc. The new system needs the same blood tests, and in practice the commonly used pre-op risk scores are easy to use and this is not a limitation in their application. The new score appears to be based on a heterogenous population,

	but with a 70% Chinese majority. Given that one of the justifications for the score is possible genetic differences, I find the claim that a score that is heavily weighted to a single race can be used for all Singaporeans of different ethnicities. These need to be addressed before I feel that this paper would be ready for publication.

VERSION 1 – AUTHOR RESPONSE

We would like to thank Dr Iain Moppett, Dr Wei Zhu and Dr Nirav Shah for their time in reviewing our manuscript and in providing their invaluable comments. We believe these comments helped to improve our manuscript greatly. Please allow us to submit a point-by-point reply in red text below.

Rev 1

1. I would like to see a STROBE checklist or similar please

Thank you for the suggestion. We have included a Tripod checklist (the checklist suggested by BMJ Open for studies deriving and validating prediction models).

2. The details of the modelling need some expansion, even if in a supplementary appendix. What were the univariate relationships (with n/N, OR, p) for each of the included variables. How were items selected or excluded? What were the selection criteria for the optimum model - AIC / BIC are probably appropriate here.

Thank you for highlighting the lack of clarity in this part of our manuscript. We have elaborated on our methodology in our manuscript. We included all variables that were clinically relevant on univariate analysis in the multivariate model to predict 30-day mortality and ICU admission greater than 24h. For the 30-day multivariate mortality model, 7 variables out of the 12 were found to be significant (table 2) after stepwise logistic regression. For the ICU multivariate model, 7 variables out of the 12 were found to be significant (table 3) after stepwise logistic regression. There were 5 common variables that were significant in both multivariate models - older age, higher surgical risk, moderate/severe anemia, ASA >= 3 and emergency surgery. Using the adjusted OR from the multivariate analysis, we assigned rank scores to each variable. We then combined the rank scores in both models to create a combined model, that can predict both mortality and ICU admission with a single set of variables (table 4). This final model has 9 variables, including Age, Surgical Risk, Anaemia, RDW, ischaemic heart disease, ASA, Surgical priority, Gender and presence of congestive heart failure. Details on 30-day mortality model development can be found on page 10; details on ICU model development can be found on page 11, and details on combined model development can be found on page 12-13. We compared the AUROC of the combined rank score model to the original OR model for each outcome and found that we did not lose a lot of predictive value when the rank score model was used. For mortality, the original 7-variable model had an AUROC of 0.931, while the combined rank score model had an AUROC of 0.934. For ICU outcome, the 7-variable model had an AUROC of 0.873 while the combined model had an AUROC of 0.863.

3. Did the authors looks for interaction between terms?

Thank you for the suggestion. We did not look at interaction terms as the objective is to derive a prediction model rather than predictors influencing the outcomes of interest (as we were not interested in the individual regression coefficients nor we are attempting to isolate the contribution of any variables without the influence of the other explanatory variables).

Shmueli G, Others. To explain or to predict? Stat Sci. Institute of Mathematical Statistics; 2010;25: 289–310.

4. I am not sure what the authors mean by a rank model. I think what they have done is (in a fairly non-linear fashion) rounded the OR to create simpler co-efficients. On the one hand I see the benefits of the rounding approach (it makes explanation to patient and clinicians easier) but it isn't needed with current online calculators (and a calculator is needed due to the logistic equation). Also, why have they chosen to decrease the impact of the high OR items?
The Rank Score model rounds off the adjusted OR to prevent giving certain variables extreme weights, eg. For 30-day mortality, in the 9-variable model, age greater than 85 had an adjusted OR of 34 but was assigned a ranked weight of 8 (see table 4). This is to prevent extreme scoring of certain parameters. We feel that we did not suffer significantly loss in the model's predictive performance. For mortality, the original 7-variable model had an AUROC of 0.931, while the combined rank score model had an AUROC of 0.934. For ICU outcome, the 7-variable model had an AUROC of 0.873 while the combined model had an AUROC of 0.863.

We will be developing an online calculator. In the meantime, an example on how to use the combined rank score model to predict mortality and ICU admission is given on page 14 line 8-20.

6. The equation for predicting mortality from the score needs to be published otherwise it is impossible to replicate (or use) the results.
We apologize for the lack of clarity in our first manuscript. Basically physicians using our model just need to use table 4 (combined rank score model) to calculate their total rank scores for each outcome. Based on their total rank scores for each outcome, they can obtain their estimate of 30-day mortality and ICU admission > 24h (table 5). We have also added a hypothetical patient as an example in our manuscript on page 15 line 6-11 to enhance clarity.

7. I am confused by the H-L plots. The apparent goodness of fit seems to be strongly influenced by the high risk decile. It would be helpful to tabulate these data. I am assuming that the plot circle areas are proportional to group size. Looking at the figures, the calibration appears to be not great at low risk. This probably doesn't matter at individual patient level (I'm not sure there is a meaningful difference between very low risk groups) but does matter if used for benchmarking.
Thank you for suggesting a H-L table to increase the clarity to our viewers. We have tabulated our H-L plots and presented this in our appendix on pages 2-10.
8. Table 3 - I'm not sure this adds much. 88% of patients are going to score 3 by virtue of age alone.
We have made significant changes to our model since the last manuscript and now, as we have decided to combine our ICU and mortality model into a single unified model that can predict both outcomes. The new model is now called the combined rank score model and it can be found on table 4. The risk categories, and their respective ascribed risk of 30-day mortality and ICU admission > 24h are shown in table 5.

9. Is there a reason why the authors didn't compare their score with SORT, POSSUM etc.? The premise of the paper is that these are unvalidated in the SE Asian population. This would be an opportunity to address that. This is important as if they don't work this needs to be shown and known about.
Thank you for pointing this out. Yes we agree that it would be insightful to see how our model performs relative to existing models such as the SORT and POSSUM. However, the variables in these models were not routinely collected in our patients, hence unavailable for analysis in our retrospective database. We have plans to collect the relevant data prospectively in subsequent studies with adequate funding. We have however compared the performance of our combined rank score model to American society of anaesthesiologist physical status score (ASA-PS), as well as an ASA propensity-matched patient sample. The AUROC figures are shown in Table 6.

10. On minor note, how were the CI for AUROC calculated (there are several ways to do this). The CI for AUROC are automatically generated in SPSS which uses this calculation:

W +/- Za*SE(W)

where W is the AUC, Za is the Z score such that (1-a) of the area under the standard normal distribution falls between -Z and Z (1.96 for a = .05, i.e. a 95% CI), and SE(W) is the standard error of the AUC.

Reviewer: 2

1. The robustness of the training and testing sets should be examined. This can be easily done by randomly choose the training and testing sets a few times, and for each pair, repeat the procedures to see whether the variables selected are stable, and whether the classification performance are comparable. Such resampling results should be summarized for its distribution in a separate section.

Thank you for the suggestion. We have reflected on this and felt that this is may not be necessary due to the large sample size of our study. We understand that resampling is frequently performed when there is a concern for sensitivity, especially if the sample size is not large enough. We believe the technique may have limited value on our large dataset, as similar results will be obtained.

Reviewer: 3

1. There are a number of grammatical and vocabulary based errors that need rectifying prior to publication.

Thank you for pointing this out. We apologize for the grammatical and vocabulary errors. We have proof-read our manuscript prior to this resubmission.

2. The reasons for the study do not appear to be clear. It is true that there is little/ no validation work for the currently used scoring systems in Asian populations, but the second reason that is dwelt upon is that these systems are unwieldy and require pre-op blood tests etc. The new system needs the same blood tests, and in practice the commonly used pre-op risk scores are easy to use and this is not a limitation in their application.

Our model only requires a single blood test, which is the full blood count (FBC) to obtain the preoperative haemoglobin level and Red Cell Distribution Width (RDW) level. The FBC is routinely performed in almost all our patients, unless they are young, male, have no medical problems and are undergoing minor surgeries. The P-POSSUM requires preoperative electrocardiogram (ECG) and serum sodium, urea and potassium, in addition to white blood cell count from the FBC. In our institution, these additional tests are only ordered for patients who are undergoing moderate to severe surgeries, or who are older and have medical problems. Thus, routine ordering of ECG and preoperative biochemistry panel for everyone in order to perform risk scoring assessment would not be cost effective for young, healthy patients undergoing minor surgeries.

Furthermore, other scoring systems have not been validated in our local population in Singapore. This is partly due to the different coding systems in every country. For example, SORT uses procedure dictionary which is country specific (based on UK's AXA PPP procedure database) and inherently different from the Singapore's Ministry of Health coding system.

3. The new score appears to be based on a heterogenous population, but with a 70% Chinese majority. Given that one of the justifications for the score is possible genetic differences, I find the claim that a score that is heavily weighted to a single race can be used for all Singaporeans of different ethnicities.

Thank you for the comment. The dataset resembles the population makeup of Singapore. Race were investigated as a possible risk factor and found not to be a significant in in our multivariate analysis, hence it was not included in the model. Differences in perioperative outcomes between regions may not just be influenced by race, but also cultural and socioeconomic factors, as well as differences in national health systems which are equally reflected on all Singaporeans.

VERSION 2 – REVIEW

REVIEWER	Wei Zhu Stony Brook University (State University of New York at Stony Brook) USA
REVIEW RETURNED	02-Dec-2017

GENERAL COMMENTS	Good revision

REVIEWER	Iain Moppett University of Nottingham UK
REVIEW RETURNED	04-Dec-2017

GENERAL COMMENTS	The authors have revised the manuscript and the statistical methods are much more clearly described - thank you. Two points for consideration. 1) I am still not completely clear about the rationale for the study. As my fellow reviewer pointed out, the study appears to be about developing / validating a Singapore specific system but there is no attempt to assess it against current models. I appreciate POSSUM may be impossible due to retrospective data collection. I think the authors could make a reasonable stab at comparing with SORT. I realise that surgical urgency is not completely classified but best / worse case imputation could address that. In essence the difference between CARES and SORT are that SORT has more surgical info, CARES slightly more physiology. 2) I am slightly confused by the new combined system. Is this predicting death AND ICU stay > 24 or death OR ICU >24? A few other points: ROC curves - this may be completely wrong on my part, but the figures don't (by eye) appear to have AUROC of 0.93. (2 and 3) whereas Appendix 1 looks much more convincing. Looking at the H-L tables (thank you for providing these) there are a lot of cells with very low expected cell counts. I am not a statistical expert but my hunch is that this should be corrected for given that H-L is just a glorified Chi-sq test. I am happy to refer such questions to someone more knowledgeable than me. Iain Moppett Nottingham, UK

VERSION 2 – AUTHOR RESPONSE

We would like to thank Dr Edward Sucksmith of BMJ Open and reviewers, Dr Iain Moppett and Dr Wei Zhu, for their time in reviewing our manuscript and providing their invaluable comments. We believe these comments helped to improve our manuscript greatly. Please allow us to submit a point-by-point reply in red text below.

Rev 1

1) I am still not completely clear about the rationale for the study. As my fellow reviewer pointed out, the study appears to be about developing / validating a Singapore specific system but there is no attempt to assess it against current models. I appreciate POSSUM may be impossible due to retrospective data collection. I think the authors could make a reasonable stab at comparing with SORT. I realise that surgical urgency is not completely classified but best / worse case imputation could address that. In essence the difference between CARES and SORT are that SORT has more surgical info, CARES slightly more physiology.

Thank you for the suggestion. We have considered and discussed extensively on the value of including a comparison with SORT in this present retrospective study. This is the first report on the CARES model, which is based on a large retrospective data of perioperative patients. There are various differences which may make a retrospective comparison suboptimal. For example, the surgical coding and classification for the various surgeries has been done differently in the CARES model compared to SORT due to differences between the local systems. In SORT, the surgical severity is divided into Minor, Intermediate, Major and X-major, while in CARES, it is divided into Minor - intermediate and major. In addition, urgency is divided in SORT into elective - expedited - urgent and immediate while in CARES it is simpler -- elective or emergency.

Furthermore, surgeries in SORT are coded with the AXA PPS coding, while we code our surgeries differently based on Singapore's national coding for procedures. This would make the task for an accurate retrospective classification and comparison extremely difficult.

In addition, we do not routinely collect malignancy status (defined by SORT as "Active malignancy within past 5 years") in our current local clinical practice, hence this variable is also not present in our retrospective dataset.

Based on all the above reasons, it may not be meaningful to do a direct comparison between the two models using our current large retrospective dataset. Nonetheless, in this study, we have compared CARES with ASA-PS classification as this is the only risk stratification model we currently use in our clinical practice in Singapore.

We fully appreciate the importance of a head to head comparison with other currently available models, as well as external validation studies of CARES. We aim to address this in future prospective work. In fact, we are currently planning a prospective study to compare CARES with SORT and POSSUM.

2) I am slightly confused by the new combined system. Is this predicting death AND ICU stay > 24 or death OR ICU >24?

A single score from the CARES model predicts both risk of death AND need for ICU stay >24 hours.

3) ROC curves - this may be completely wrong on my part, but the figures don't (by eye) appear to have AUROC of 0.93. (2 and 3) whereas Appendix 1 looks much more convincing.

Thank you Dr Moppett for your close scrutiny and sharp eyes. We thank you for the chance to correct our mistake. We have ran multiple analyses and accidentally pasted the wrong ROC curves. We have double-checked with all parties involved and clarified and have replaced the figures with the correct ROC curves in this second revision.

4) Looking at the H-L tables (thank you for providing these) there are a lot of cells with very low expected cell counts. I am not a statistical expert but my hunch is that this should be corrected for given that H-L is just a glorified Chi-sq test.
Thank you for pointing this out. We have consulted our statistician, Dr Chan Yiong Huak, who is the Head of Biostatistics Unit in Yong Loo Lin School of Medicine, National University of Singapore on this matter. He is also one of the co-authors in this paper and performed the H-L test. The fisher's exact test should be used when the expected frequency is less than 5. However, by looking at the obtained chi-square p-values, the differences between the fisher's exact p-value and chi-square p-value would be very small and hence the chi-square p-values should be acceptable. None of the p-values tabulated for the combined model (appendix Va to Vd) are approaching significance level ($p<0.05$).

VERSION 3 – REVIEW

REVIEWER	Iain Moppett University of Nottingham. UK
REVIEW RETURNED	22-Jan-2018

GENERAL COMMENTS	The authors have provided good reasons for disagreeing with my previous suggestions and have corrected the errors in the figures.