Author’s response to reviews

Title: Validity of the CR-POSSUM model in surgery for colorectal cancer in Spain (CCR-CARESS Study) and comparison with other models to predict operative mortality

Authors:

Marisa Baré (mbare@tauli.cat)
Manuel Jesús Alcantara (malcantara@tauli.cat)
Maria José Gil (mgil@parcdesalutmar.cat)
Pablo Collera (pcollera@althaia.cat)
Marina pont (mpont@tauli.cat)
Antonio Escobar (antonio.escobarmartinez@osakidetza.net)
Cristina Sarasqueta (CRISTINA.SARASQUETAIZAGUIRRE@osakidetza.net)
Maximino Redondo (mredondo@hcs.es)
Eduardo Briones (eduardo.briones.sspa@juntadeandalucia.es)
Paula Dujovne (pdujovne@fhalcorcon.es)
Jose Maria Quintana (josemaria.quintanalopez@osakidetza.net)

Version: 1 Date: 11 Jan 2017

Author’s response to reviews:

We are grateful to the reviewers, who demonstrate their knowledge of the methodology of this kind of studies and make some very interesting contributions. Before responding to the reviewers’ comments point by point, we think that we need to clarify some general points about our study to respond to some of the reviewers’ observations and comments.

This manuscript derives from a coordinated study that started in 2009. Among other objectives, the study sought to validate the CR-POSSUM model in the Spanish context, because some departments of surgery in our region used this model for clinical purposes although its validity had not been evaluated in our environment. The variables were recorded during the study have enabled us to validate not only the early versions of the POSSUM but also other models that have been developed in recent years in our international context (Europe), since we had collected data on all the factors that they consider. This is the reason why we decided to also validate the
IRCS and AFC models that had been developed and published before the analyses in this manuscript had finished.

The purpose of this project is to improve clinical practice through the use of instruments that help decision making at the bedside. Therefore, our first objective is to identify the most common methods or models in our context, to analyze the appropriateness of their use, and to develop better models for practical use.

Please see our point-by-point responses to the reviewers here below.

Reviewer #1: This is a well-written research paper to estimate the operative mortality in surgery for colorectal cancer in Spain, to validate/recalibrate the CR-POSSUM model, and to compare its discriminatory capacity with other models. The overall evaluation in methods and approach seems to be appropriate but shows shortage and limitation in describing methods section and in providing sufficient and persuasive information for the readers.

Background:

* Page 6 Line 113-114: Elaborate or give example of the over/underestimation of risk. I think some of the models overestimate the risk of mortality in low-risk patients. When does underestimation occur?

We agree with you that these models tend to overestimate risk, especially in low risk patients, as we mention in the background section. On the other hand, only few authors (not cited in our paper) have found some underestimation in some low-risk patients with the P-POSSUM model (Merad F, et al. World J Surg. 2012; Tamijmarane A et al. World J Surg Oncol. 2008). Because it is not so frequent, we have decided to avoid going deeper into the question of this model underestimating the risk.

* Page 6 Line 115: Is this model used by AFC a validated scoring system? What was the rationale for including this model?

The AFC model has been validated in various studies published between 2007 and 2013. The rationale for including this model is explained in the general comments above. We hope that this explanation clarifies why we have included some models in addition to the CR-POSSUM.
Page 6 Line 125-126: What interventions - postoperative or preoperative? Is this to say that presence/absence of interventions may affect patient outcome? Or is the surgery the intervention? Please elaborate.

Thanks for pointing out this ambiguity. We have added the word 'surgical' in the text in order to clarify that we refer to surgical interventions (or surgery).

Methods & Results:

Page 7 Line 135-140: What was the composition of the type of hospitals - rural/urban, small/large etc.?

We agree that this is important. We have added a new reference about the first article of this project in which we clearly explain more details about the methodology. In fact, different types of hospitals participated in the project: rural and urban, large and small, teaching and non-teaching hospitals.

Page 9 lines 182-192 Models for predicting the risk of death: what about the P-POSSUM model? Authors need to explain how to construct a physiological score and a operative severity score for three models of POSSUM, P-POSSUM, and CR-POSSUM as well as general information about those scores such as range, how many questions, interpretation based on scores, etc.

When we refer to POSSUM models we meant POSSUM and P-POSSUM. To make it clearer which models we refer to, we have added the P-POSSUM in the sentence.

We have also added an explanation to the paragraph and more information in the supplementary material to show how the score is calculated and how it is used. The scores do not have a direct clinical interpretation per se; rather the authors of the scores introduced them directly in the formula, representing the sum of various clinical variables.

Page 11 Lines 233-234: There is no enough information about how each predictor variables to be recalibrated (authors need to provide a table about this information for readers).

To clarify this point, we have added a brief explanation about the calculations involved in recalibrating the model in the methods section.
Overall comments in methods: IRCS model is different from the other three models of POSSUM, P-POSSUM, and CR-POSSUM on constructing prediction models. Do you think that we can compare IRCS model compares to the other three? What are common variables among four models? What variable was used as a base variable to construct weight system? How did you control multicollinearity issue among similar variables?

The IRCS model uses all the same variables as the other models. In fact, what changes are the categories that are defined. Like the other models, the IRCS model also uses logistic regression to adjust estimators for each variable that remains in the final model and defines a score based on the scores assigned to each category of variable. One aspect that makes the IRCS different from the other models is that it uses variables known only just before entry into the operating room, making it possible to estimate the probability of the patient dying, which is important for decision making. In fact, the percentage of missing variables for this model was very low, showing that at least it is easy to compile the variables in the model.

In relation to the interesting points that you bring up with regard to the weights and multicollinearity, we consider that these aspects are very important in the process of developing predictive models and that the authors of the different models we use would best be able to respond to these questions. Our aim was merely to validate or recalibrate the models, and as such we did not consider re-estimating the weighting or revising the collinearity of the variables.

Discussion:

* Page 12 Line 258: Why do you think the operative mortality in your sample of patients was also lower compared to other studies? As mentioned above, perhaps looking at the type of hospitals urban/rural or available technologies/skilled staff would be of interest to see if any of your results may be explained by these factors?

This is an important observation, since the low mortality in our study makes it much more difficult to demonstrate the discriminative capacity of the models. We think that our study was conducted in a period where mortality was lower than in previous studies, very possibly due to improvements in the surgical management of patients in general and perhaps also to the selection of candidates for surgery. Given that the study includes a wide variety of hospitals that is representative (as is explained in the new reference that explains these details.BMC Cancer), we do not think that the factors you point out are the reason for our low mortality (range 0% – 3.6%).

* Page 12 Line 264: please give examples for "… due to various factors"

We have added some examples.
Please explain further what the nature of the missing data was (not collected, incomplete record, etc.) and how the analysis was conducted (listwise deletion or pairwise deletion).

The main reason for missing data in some of the variables is that they are not determined routinely and, thus, it is impossible to collect them (e.g., urea, total blood loss).

With respect to the analysis, we did not assign values to any of the missing variables—patients for whom one of the variables in the model was missing were directly eliminated from the sample analyzed (listwise). Table 3 shows that sample analyzed is different for each model validated, and this depended on the information missing for the variables included in the different models. Thus, we were able to use 77% of the sample for the POSSUM models, 83.4% for the CR-POSSUM, only 53.3% for the AFC model because of the absence of the weight in the clinical histories, and 96.5% for the IRCS model.

In-hospital mortality AUS were similar for POSSUM, P-POSSUM, CR-POSSUM, and IRCS. Can you really say that IRCS had the greatest discriminatory power when the difference is less than 1 with wide confidence intervals?

Thanks for this comment. As you point out, we cannot really say that. Instead, we have softened the statement and have changed the verb.

Reviewer #2: This is a well written paper with clear methods and results. However, the authors do not make sufficient case that the study is needed. Part of the issue with this is that they only compare 4 models and there are many in the literature, some of which they reference in the discussion. Justification is needed for the choice of these 4 models. Secondly, there are already external validation studies comparing the models in similar patients, as referenced by the authors. Thirdly, the authors are not clear whether they are comparing a model for risk adjustment in order to make comparisons between healthcare providers or a prognostic model to guide clinical decision making. This is an important distinction which affects the requirements of the model and the variables that should be included in it. Fourthly they do not sell this comparison of models as an external validation or make wider recommendations from their findings about which model should be used. Instead they focus on the use of the models in one setting (Spain), which limits the potential of the study.
The main conclusion of the study is that a new model is needed to discriminate patients with high risk of death. This conclusion cannot be drawn when only 4 models have been compared.

We understand your concerns and hope that the explanation provided in the introductory comments helps to clarify the aims of our study and selection of models.

The purpose of our study was mainly to improve clinical practice, and in fact most of the models published to date were developed with similar objectives in mind. In our environment, clinical departments are most familiar with the POSSUM models or their derivatives and that is why we considered this an appropriate starting point for our research line. We agree that our approach does not allow us to recommend one model over another on; rather we hope to arrive at a simple, validated, feasible instrument that clearly distinguishes the patients who have the greatest surgical risk in surgery for colon surgery.

Some more specific comments:

1. Line 201: R is not defined. It is not necessary to give the link function for a logistic regression as this is standard.

Thanks for this observation. We have eliminated the link function.

2. The "enter" method of recalibration needs more explanation and a reference. Does this method differ from re-fitting the model with the same variables in the new dataset?

Many thanks for this comment. We agree with you that the term 'enter' is not necessary. Instead, we explain that standard regression techniques are used, which means the same as re-fitting.

3. What was done about missing data?

The main reason for missing data in some of the variables is that they are not determined routinely and, thus, it is impossible to collect them (e.g., urea, total blood loss).

With respect to the analysis, we did not assign values to any of the missing variables—patients for whom one of the variables in the model was missing were directly eliminated from the sample analyzed.

4. In the results it needs to be clearer whether each of the AUCs is for re-calibrated models.
Thanks for this observation. We have made this clearer.

5. When looking at calibration of the models, more emphasis should be given to the calibration plots and less to the statistical test (Hosmer-Lemeshow) as the latter is under-powered to give evidence of a lack of calibration with fewer than 50 deaths.

We agree with this comment, and changed the sentence.

6. The impact of exclusions and missing data is not sufficiently acknowledged as a limitation, or investigated where possible.

Following your suggestion, we have provided a more thorough explanation of the consequences of missing data for some of the variables in the models. As in all observational studies that do not interfere in routine clinical practice, missing values can be a limitation. In our study, although a small percentage of the missing values can be attributable to data collection errors, in most cases missing values are more related to the fact that the test that would have provided the missing values was not done.

7. The authors describe the study as large and yet there were fewer than 50 deaths. This needs to be acknowledged as a limitation.

We agree that this is a limitation and mention it as such. We have strengthened the statement: "On the other hand, the mortality rate was low, with fewer than 50 deaths in both mortality indicators, and this might have compromised our capacity for recalibrating the models".

8. The authors say that none of the models they compare is specific to colorectal cancer patients. But at least two of their references are models specific to these patients they need to justify why these models were not included in their comparisons.

This is a very relevant observation. However, there are reasons that justify why these models could not be included in our study. For example, Walker et al’s model was published after we had finished our analysis and prepared our manuscript; De Vries et al’s model was elaborated outside the European context and also uses administrative and census variables that were unavailable to us and outside the scope of our study.