OFFICIAL

The Editor,

The Baishideng Publishing Group (BPG)

Dear Editor:

Thank you for giving us the opportunity to respond to the valuable suggestions and comments made by the reviewers of our manuscript entitled Preoperative Risk Modelling for Oesophagectomy – A Systematic Review (Manuscript number 80597). We have carefully considered each of the issues raised by the reviewers and addressed each comment in a step-by-step fashion below.

These changes have been reflected in the manuscript submission as well. In addition to this, in the course of combing over the manuscript in great detail, the authors noted a few small errors which they have corrected themselves. These are also outlined below.

Thank you for your consideration of our reviewed manuscript in advance and we await eagerly the response of the reviewers and editorial board.

Sincerely,

Dr James Grantham
Department of General Surgery, Modbury Hospital
Modbury, South Australia, 5092
Phone: 0447980425
Email: Jamespgrantham91@gmail.com

Reviewer 1 Comments with Responses:

1. **Abstract/Methods: The article refers to conducting the review in line with PRISMA, however this is a reporting tool for the subsequent write up, not a guide for the review conduct – please rephrase accordingly.**

This has been corrected:

In the method section of the abstract: 'The articles generated from each search were collated, processed and then reported in accordance with PRISMA guidelines'.

In the materials and methods section: 'The articles generated from each search were collated and processed with reporting in accordance with the PRISMA'

2. **The article refers to PRISMA 2009 guidelines, however these were updated in 2020, could the reporting checklist please be updated accordingly? This may have minor considerations for e.g. Figure 1 layout.**

The World Journal of Gastrointestinal Surgery requires submitting authors to satisfy the PRISMA 2009 checklist statement which is why we have left it as it stands on the first page above the abstract in the manuscript. We note the reviewers point regarding the updated PRISMA guidelines and have reported our search strategy in the updated format, referencing the 2020 guidelines now.

3. **The systematic review did not follow a pre-published protocol, and the first three sections of the methods (search strategy/article selection, inclusion and exclusion criteria, and data extraction and synthesis) are difficult to follow in places as a result. Were 'PICO' criteria designed for the inclusion/exclusion criteria? Some vague statements are included, for example:**

a. **Articles that exclusively assessed distant outcomes….were excluded' – what was considered a distant/long-term survival outcome? Were peri-operative mortality outcomes considered to be e.g. 30 or 90 day mortality rates?**

We have elaborated on our definition of perioperative period the inclusion criteria by adding:

'The perioperative period was defined as any duration whilst an inpatient from the index oesophagectomy admission and no more than 90 days post-operative if the patient had been discharged'

b. Studies which presented insufficient data for meaningful analysis were excluded – what data were required to be deemed meaningful analysis?

We have elaborated on this as follows:

'Studies which presented insufficient data for meaningful analysis, such as calibration measures in the form or p-values or area under the receiver operating characteristic curve (AUC) and/or discrimination statistics, were also excluded'

c. Non-English language studies were excluded, however in the results section 'Methodological quality – study participation', the first sentence refers to a prognostic nutritional index study being unavailable in English – surely this should have been excluded according to the inclusion/exclusion criteria?

Within the 27 studies which were included, 13 were articles that developed new models and 14 were articles that externally validated existing models. All of these articles were fully accessible in English. The article which appraised the PNI model was published by Filip et. al. in 2015 (This article also served as the development study for the PNI-Multivariate model and external validation of multiple other models including the Charlson Comorbidity Index, Age-Adjusted Charlson Comorbidity Index, POSSUM and the Amsterdam Score) and was eligible for inclusion as it was published in English. There was a total of 21 models appraised with 13 having their development models captured amongst the 27 studies as mentioned above. We looked at the index articles which developed the remaining eight models which had been externally validated amongst the 27 articles captured by the systematic review. Some of these included the original article in which the POSSUM score was developed, the original article in which the Charlson Comorbidity Index was developed etc. This was done so that we could appraise the methodological quality of how these models were developed. One of the eight original articles was not available in English (Onodera 1984) and hence why it was not available for appraisal of methodological quality. This

article, along with the seven other articles in which these models were developed were not part of the 27 captured in the systematic review process and hence whether they were fully available in English was not considered.

d. **What subgroup analyses were planned versus those conducted? There is reference in the methods to planned analysis by heterogeneity of surgical method, but I did not see much in the results text on this. Vice versa, the results text describes the number of studies by histological subtype of oesophageal cancer – this was not mentioned in the methods.**

I have amended the text to detail that subgroup analysis by surgical method was not possible due to multiple approaches being incorporated in many of the studies. I have also mentioned the same limitation existed for subgroup analysis in terms of histological subtype and neoadjuvant chemotherapy status. There was a mention in the submitted manuscript that we were collecting the histological subtype and neoadjuvant therapy status "Patient characteristics including the proportion of neoadjuvant therapy use and histological subtype were also extracted."

e. **In the results text, methodological quality section, a sentence refers to one development study including fewer than 100 patients – was this sample size a required criteria for inclusion or quality assessment?**

No it was not an inclusion criteria. It was a component of the quality assessment framework we utilised to appraise methodological quality of a development study for a multivariate risk model. This protocol was developed by Minnie et. al. as mentioned in the manuscript and has been applied in many studies assessing multivariate risk models.

f. **In Figure 1, abstracts superseded by articles were excluded. Were conference abstracts included if subsequent full text articles hadn't been published? Overall, this section would benefit from greater clarity as it would not be**

possible for a researcher to fully replicate the systematic review methods in future without more precise definitions here.

We have added to this section to further clarify the issue, it now reads:

Abstracts that were superseded by full articles were excluded. Abstracts from conference proceedings not subsequently published in full were considered eligible for inclusion, provided it included sufficient data for meaningful analysis as outlined above.

4. **It would be helpful to include reference numbers throughout the results section and descriptive tables when referring to studies, e.g. in the clinical credibility section 'six models scored the highest….' – it is not immediately obvious which publications these relate to.**

We have incorporated references to each of the tables for each study referenced. We have also included in-text reference numbers to any situation where a specific study or model is referenced in the results and discussion section. There has been a conscious decision not to incorporate references into the text where there is a large number or models referred to for two reasons. The first is that it is much easier to refer to the tables to identify specifically which models scored what for each element. The second reason is that it would make the text too cumbersome with huge numbers of in-text references trailing at the end of every sentence.

Below is an example of a paragraph if we were to have incorporated such references as suggested:

The median clinical credibility score, out of 7, was 5.5 (range 4.5-6) (Table 3). Six models scored highest at 6 out of 7: the Rotterdam, Philadelphia, Amsterdam, PNI, and the original and revised STS models[30-33, 37, 56]. Twelve of these twenty-one preoperative models were oesophageal-specific and all models provided timely data for clinical decision making[29-34, 35-38, 39-40]. Three of these models used subjectively reported patient health questionnaire data[41, 61, 63]. Seventeen of the twenty-one preoperative models were considered easy to generate with the other four reliant on pre-operative spirometry, which may not be routinely performed[29, 34, 38]. Three of the 21 preoperative models were considered challenging to understand[35-36, 40]. Sixteen of

the twenty-one preoperative models were found to generate a useful scoring range to prognosticate patient outcomes[30-37, 40-41, 56-58, 60-61, 63].

5. **Results section – Study characteristics: 'The histological subtype of oesophageal cancer was reported in 16 studies, including six from Asia and 13 from Western nations' – 6+13=19 studies? This may be a discrepancy in the number of models versus publications, some clarification would be helpful.**

This was an error which has been addressed. It arose as 3/6 studies originating in Asia reported histological subtype. We accidentally put the six figure in instead of the three. It has been amended to: 'The histological subtype of oesophageal cancer was reported in 16 studies, including three from Asia and 13 from Western nations'

6. **The presentation of results tables could be enhanced, with consideration given to presenting separate tables by the primary and secondary outcome measures (this would help to avoid overly long tables) and footnotes are needed to explain the many abbreviations within the tables. More cross-references to tables and figures in the results text would also help to guide the reader.**

This is a good point and the authors agree with the suggestions. We have separated the model performance into four separate tables (Now Tables 5-8). These tables address model performance in predicting mortality, major morbidity, overall morbidity and respiratory complications/return to theatre/readmission/anastomotic leak outcomes respectively. We have added further footnotes to guide the use of the abbreviations present within the tables. We have also added some more cross-references in the results section accordingly. In terms of the abbreviated model names in the tables, this would lead a very large number of footnotes. As such, we have introduced a new figure, labelled figure 2 as a reference key for all of the model names. We have also attempted to clean up the formatting of the tables in terms of lines and spacing in the hope that they appear sharper now.

7. **A quality assessment/risk of bias tool needs to be applied to all included studies to determine 'high-quality' models, rather than high-quality models being determined solely on the basis of the resulting AUC.**

We agree that this is essential. This is the methodological quality component of the study. We adopted the quality assessment framework of Minnie et. al. This is a framework with 20 criteria with well performing models being of strong methodological quality and thereby having a low risk of bias in their development. The models' performance against this framework is discussed extensively throughout the results and discussion section and summarised in Table 4. This methodological quality along with model performance (discrimination/calibration), clinical credibility, external validation status and clinical effectiveness formed the five components of how we appraised the each model (As shown in Table 9).

8. **A cut-off of 0.70 is described as determining clinical usefulness – is this based on published literature/a widely accepted figure?**

There is literature available which advocates for other cut-offs. In arriving at this figure, we referred to various sources to determine which is the most widely accepted threshold is. We arrived at this figure based on our reading, with several articles advising 0.7 to be a widely or conventionally accepted threshold for adequate discrimination. This was based on several sources, including Applied Logistic Regression by Hosmer and Lemeshow, among others. Here are some references supplied.

Li F, He H. Assessing the accuracy of diagnostic tests. *Shanghai Arch Psychiatry*, 2018 June 25; 30(3):207-212. [PMID: 30858674. PMCID: PMC6410404. DOI: 10.11919/j.jssn.1002-0829.218052].

Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010 Sep 1;5(9):1315-6. [PMID: 20736804 DOI: 10.1097/JTO.0b013e3181ec173d]

Hosmer DJ, Lemeshow S, Strurdivant RX. Applied Logistic Regression, Third Edition. Hoboken, NJ: John Wiley & Sons, Inc.; 2013.

We have changed the reference in the text to reflect Hosmer and Lemeshow, rather than a different article, as we believe it carries more weight on the matter. Acknowledging that there are proponents for different thresholds such as 0.75 and 0.8, we have softened the language around the 0.7 threshold.

9. **Through the manuscript, particularly the results and discussion, suggest to use the phrasing peri-operative or short-term mortality outcomes rather than general mortality terms, given the inclusion/exclusion criteria above.**

In accordance with the advice of the reviewer, we have added perioperative to the headings of all of the major result headings. So whereas it previously stated 'Model Performance – Mortality', it now reads 'Model Performance - Perioperative Mortality'. Each table description also specifies it is perioperative outcomes reported. We have incorporated the term perioperative at various points throughout key stages including the abstract, results and discussion section. It is not specified in each instance when referring to mortality that it is perioperative mortality being considered as we believe with sufficient specification woven through the text, it is broadly implied in these areas and too cumbersome within the text to precede each reference to mortality as perioperative mortality. For similar reasons within the tables, we have left the column relating to the outcome measured as mortality, major morbidity, morbidity etc (As adding perioperative in there would be make the whole column much larger and messier in appearance, with the overall table much larger as a result).

Reviewer 2:

1. There are quite a lot of space errors in the text of the article I uploaded (lack of spaces leads to the merging of two or more words into one), which makes it difficult to read and is unacceptable in the final version of the article.

We did not have an issue that we were aware of in the manuscript we submitted. We have gone through the article to ensure consistent and appropriate spacing between both text and line spacing in general.

2. There are a plethora of single factor prognostic indicators (Page 9) - maybe, "a real plethora"?

The authors feel it reads better as presently written.

3. The principles of formation of tables 2-5 are not quite clear. I got the feeling that the various predictive models are randomly placed in them. Whereas for a scientific article it is necessary to follow some order: alphabetical, chronological, by the number of cases studied, etc. If my assumption is right, then it seems to me correct to change the order of the lines in these tables.

We agree with this statement. We have now organised the tables a little better. Tables 1 and 2 and ordered by publication date of each article in chronological order. Table 3 is ordered in publication date of the original development study which led to the creation of each model (Such as PNI being from Onodera 1984, CCI being from Charlson 1987, ACCI being from Charlson 1994 and so on). This order is carried through in Table 4. Tables 5-8 have only some models included in each table where we have maintained this established order for the sake of consistency. All models are summarised in table 9 in this order also.

Additional Revisions Made by the Author:

In the course of reviewing the manuscript in great detail once again, the authors have noted a few small errors that, whilst not identified as revisions requested, ought to be corrected so that the manuscript is of the highest quality. This has resulted in a small number of changes that were not requested.

First of all, the PNI multivariate model which was devised by Filip et al. in 2015 was noted to be predicting major morbidity, rather than overall morbidity. The confusion seems to have arisen as this article also validates a number of existing models against morbidity outcomes then goes on to develop the PNI multivariate to specifically test it for major morbidity. The article took quite a few read throughs before this became evident.

Similar to this article by Filip in 2015 which developed a new risk model whilst validating other models, Wan et. al. did the same in 2020 when they devised the RAI-Revised (Cancer Corrected) model and simultaneously externally validated a number of other models in the same paper. It stands to reason that is article be included in the development study list (As Filip et. al. has been). Therefore with this reclassification, the list of development studies has increased from 12 to 13 whilst the number of validation studies has reduced to 15 to 14. The number of models that were considered to have been externally validated also reduced by one, from 15 to 14.