PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A Systematic Review of Methodological Quality of Model Development Studies Predicting Prognostic Outcome For Resectable Pancreatic Cancer
AUTHORS	Bradley, Alison; Van Der Meer, Robert; McKay, Colin

VERSION 1 - REVIEW

REVIEWER	Kurinchi Gurusamy University College London, UK
REVIEW RETURNED	05-Nov-2018

GENERAL COMMENTS	• Title is not reflective of the objectives: The current title suggests that they are looking for the performance of the different models rather than the methodological quality. • The authors have restricted the language to English. This is mentioned as a limitation in the strengths and limitation section. There is no further discussion on this issue or attempts to reduce this. Therefore, the systematic review is biased. If the authors cannot address this limitation, then they should call this a scoping systematic review. • How many people screened the data? • “Search design and data extraction was performed by the lead reviewer and with second author performing independent quality assurance”: This is unclear. Was this independent data extraction or independently checking the data? Was this done for all studies or a sample of studies? • There is too much discussion about Bayesian multivariate analysis and the article seems like an advertisement for Bayesian multivariate analysis – there is no evidence from the studies that they reviewed that Bayesian multivariate analysis provided better prognostic results than other ways of predicting and all the discussion about the advantages of Bayesian multivariate analysis is inappropriate and should be removed. • While this is mentioned as a narrative systematic review, it will benefit from a table detailing the classification of the domains and the reasons for the classification. At present, it is difficult to follow.

REVIEWER	Dr Luke Hodgson Western Sussex Hospitals NHS FT, UK
REVIEW RETURNED	13-Nov-2018

GENERAL COMMENTS	Many thanks for the article which I enjoyed reading. This is an important area where prediction models could potentially very useful . As the authors point out there have been an increasing number of published models recently. The paper uses appropriate methodolgy to critique the papers however I have some areas to address: 1. Make clear to the reader EPV is usually quoted as number of predictors assessed compared to number of events ie for these models, deaths. There has been some recent research suggesting EPV of 10 may be too simplistic see: Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2018 Oct 24. doi:10.1002/sim.7992. 2. Could you more clearly summarise in a table/figure the most frequently included variables? – as most models have significant shortcomings other researchers may want to investigate these 3. Most papers use univariate analysis to select for multivariate analysis – not recommended eg in TRIPOD guidance 4. Internal validation rarely done: 3 bootstrapping and 2 random split 5. External validation: you say references 17-20 but I don’t think reference 20 has an external validation? Furthermore the others have issues: Reference 17 used 17 variables and the external validation had only 61 patients; Reference 18 – used univariate to select for multivariate analysis with 56 variables and the outcome in 78 (derivation) & n=43 (External validation). Reference 19 – Hodson spelling correction – in the paper unclear how many events occurred in the external validation cohort - no papers described to have externally validated one of the models elsewhere / separate to the derivation authors, is a significant limitation that should be perhaps suggested as a future avenue of research? 6. Calibration – this is crucial & needs further emphasis in the discussion that it was missing frequently or not performed adequately eg the Xu paper where their calibration curves came from the derivation data; in this paper why is there such a difference between AUC and C-Statistics? 7. No mention of Impact analysis - the next step after external validation 8. A note could be made that in the 2 studies using alternate methodology AUCs were not impressive: Walczak – neural network AUC 0.66, Smith & Mezhir Bayesian model – AUC 0.65. 9. Consider whether Risk of bias could be performed – PROBAST http://s371539711.initial-website.co.uk/probast/ – the author of this can be contacted for permission to use this tool, yet to be published: Robert@systematic-reviews.com 8. IPD meta-analysis is another avenue of future systematic reviews you could briefly mention in the discussion: Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140 doi:10.1136/bmj.i3140

	9. In the discussion there is a rather long explanation of Bayesian networks that could be considered outside the scope of the article or shortened? 10. In the discussion perhaps you could emphasise the limited discriminatory performance of most models (usually AUC <0.7) – would such discrimination outperform clinical opinion / be useful? Those with an AUC approaching 0.9 are on small sample sizes with probably overfitting.

REVIEWER	Johanna Damen University Medical Center Utrecht
REVIEW RETURNED	12-Feb-2019

GENERAL COMMENTS	Predictive Prognostic Modeling For Resectable Pancreatic Cancer: A Systematic Narrative Review bmjopen-2018-027192 This article describes a systematic review of prognostic models for predicting post resection prognosis in patients with pancreatic ductal adenocarcinoma. The quality of the existing models for these patients has been described. Although I do think the authors address a potentially relevant topic, I have several concerns: Major concerns: - What is the actual focus and aim of the review? Is it to describe the quality of the available models? Is it to teach about best practice when developing prediction models? Or is it to convince the reader that Bayesian Networks are a better method for making a prognosis? I believe these are three different papers and I suggest the authors choose one of these aims. - From the introduction it is not clear how prediction models can be used in clinical practice. Which decisions can be made using these models for this specific disease? Please specify this. - The search strategy is not fully reported. Please adjust Supplementary Material S1 to make sure the full strategy is listed, including all Boolean operators, as it is now unclear how search terms were combined. Further, there are duplicate terms in the list currently provided (e.g. ‘model’, ‘models biological’ and ‘models biologic’ are listed more than once). It is also not clear whether Mesh/Emtree terms have been used or free text. Please also add this. At this moment I do not have enough information to judge whether the search is appropriately conducted. - Why are studies that externally validate a model without updating the model excluded? I believe these types of studies can provide most valueble information about the quality and usefulness of currently available models and therefore should not be excluded. - There are several mistakes in the flow chart (Figure 1). How is it possible that the number of records screened after duplicates removed is lower than the number of records identified in Embase? Usually this is never lower. Also the difference between the number of records identified using PubMed/Medline is very low compared to the number identified with Embase. Usually there is a difference indeed, but I have never seen a difference this large. Some of the numbers are also listed in the wrong box: the number of records screened should be 23,097 and not 263, and the number of full-text articles assessed for eligibility should be 263 and not 15. The authors can also remove the last box (studies

	included in quantitative synthesis) as a meta-analysis was not performed. - The conclusion is not based on the data presented in the results. Minor concerns: - There is some strange terminology used. What is a narrative systematic review? I think this is a systematic review, as it follows a systematic approach for identifying studies and collecting information. I do not see where the review is narrative. Further, with predictive we usually refer to factors or models predicting response to treatment, see for example PMID: 19383314 or the Cochrane guidance (https://methods.cochrane.org/prognosis/our-publications). To prevent confusion, it would be better to mention it as prognostic models or prediction models. Further, it is written that the authors ‘included only prognostic multivariable prediction studies where the aim was to identify a causal relationship between two or more independent variables and the outcome of prognosis (-should this be “interest”?-), to predict prognosis’. In the field of prognosis we are by definition not looking for a causal relationship. Yellow fingers can be a perfect predictor for lung cancer, but this is of course not causally related. Please remove the word causal here. Please also change univariate and multivariate to univariable and multivariable, because I believe this is what is meant by the authors. - A review cannot follow the PRISMA checklist, it is reported according to the PRISMA checklist (page 7, line 9) - Page 10, lines 12-28, this information could be better structured below the other headings, to prevent duplicate information. - Figure 2 is better readable as a bar chart. The x-axis has a strange numbering, and it now reads like in 2005, 2007 and 2008 also 1 publication was found, while I assume this was not the case (?). And why are 2017/2018 combined? I would split this up to be able to compare 2017 to the other years.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1: Kurinchi Gurusamy

Thank you kindly for your insightful review of our paper. We appreciate your comments and have responded to each one to the betterment of the paper for which we thank you. Responses to comments:

•Title is not reflective of the objectives: The current title suggests that they are looking for the performance of the different models rather than the methodological quality.

Response: Title has been amended to reflect focus on review of methodological quality

•The authors have restricted the language to English. This is mentioned as a limitation in the strengths and limitation section. There is no further discussion on this issue or attempts to reduce this. Therefore, the systematic review is biased. If the authors cannot address this limitation, then they should call this a scoping systematic review.

Response: in the methods section it is now explained that the initial title review was restricted to English language but steps are detailed as to how full papers not available in English language were translated

•How many people screened the data? “Search design and data extraction was performed by the lead reviewer and with second author performing independent quality assurance”: This is unclear. Was this independent data extraction or independently checking the data? Was this done for all studies or a sample of studies?

Response this is now made clear in the methods section. "Search design and data extraction was performed by the lead reviewer and with second author performing independent data checking on all studies.•

There is too much discussion about Bayesian multivariate analysis and the article seems like an advertisement for Bayesian multivariate analysis – there is no evidence from the studies that they reviewed that Bayesian multivariate analysis provided better prognostic results than other ways of predicting and all the discussion about the advantages of Bayesian multivariate analysis is inappropriate and should be removed.

Response: This section has now been removed

• While this is mentioned as a narrative systematic review, it will benefit from a table detailing the classification of the domains and the reasons for the classification. At present, it is difficult to follow.

Response: this information is now provided in table 1

Reviewer: 2 Dr Luke Hodgson

We would like to thank the reviewer for his thoughtful review of our paper. We have responded to each of his comments and as a result we feel this has improved our paper thanks to his time, effort, and insight which is very much appreciated by all the authors.

Responses to comments as follows:

1. Make clear to the reader EPV is usually quoted as number of predictors assessed compared to number of events ie for these models, deaths. There has been some recent research suggesting EPV of 10 may be too simplistic see:

Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2018 Oct 24. doi:10.1002/sim.7992.

This has now been made clear in the subsection 'Statistical Power: sample size and missing data' of the results section with reference to the the recommended paper.

2. Could you more clearly summarise in a table/figure the most frequently included variables? – as most models have significant shortcomings other researchers may want to investigate these

Response: These are now summarised in Figure 3 and table 2

3. Most papers use univariate analysis to select for multivariate analysis – not recommended eg in TRIPOD guidance

Response: this fact has been more clearly highlighted in the subsection 'Model development' with specific reference to the TRIPOD guidelines

4. Internal validation rarely done: 3 bootstrapping and 2 random split

Response: this key point has been emphasised in the 'Model performance and evaluation' subsection of the results section and again has been discussed in the discussion section.

5. External validation: you say references 17-20 but I don't think reference 20 has an external validation? Furthermore the others have issues:

Reference 17 used 17 variables and the external validation had only 61 patients; Reference 18 – used univariate to select for multivariate analysis with 56 variables and the outcome in 78 (derivation) & n=43 (External validation).

Reference 19 – Hodson spelling correction – in the paper unclear how many events occurred in the external validation cohort

- no papers described to have externally validated one of the models elsewhere / separate to the derivation authors, is a significant limitation that should be perhaps suggested as a future avenue of research?

Response: You are correct that reference 20 did not undergo external validation and this has been corrected. The points you correctly raise regarding references 17 to 19 have been included in the subsection 'Model performance and validation'. Furthermore, spelling error for reference 19 has been corrected. (I refer to original reference number which have now changed in the revised version as additional references were added during the revision process)

6. Calibration – this is crucial & needs further emphasis in the discussion that it was missing frequently or not performed adequately eg the Xu paper where their calibration curves came from the derivation data

Response: The importance of calibration has been emphasised in both the results and discussion sections with issues such as those correctly highlighted in the paper by Xu et al highlighted to emphasise the issue.

7. No mention of Impact analysis - the next step after external validation

Response: now included in discussion section

8. A note could be made that in the 2 studies using alternate methodology AUCs were not impressive: Walczak – neural network AUC 0.66, Smith & Mezhir Bayesian model – AUC 0.65.

Response: now included in the discussion section

9. Consider whether Risk of bias could be performed – PROBAST http://s371539711.initial-website.co.uk/probast/ – the author of this can be contacted for permission to use this tool, yet to be published:

Robert@systematic-reviews.com

Response: ROB now included in methods section and in supplementary material S3 with reference to the recommended paper, thank you.

8. IPD meta-analysis is another avenue of future systematic reviews you could briefly mention in the discussion: Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big

datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140 doi:10.1136/bmj.i3140

Response: now included in discussion section when discussing future direction of research again with reference to the recommended paper.

9. In the discussion there is a rather long explanation of Bayesian networks that could be considered outside the scope of the article or shortened?

Response: this has been removed

10. In the discussion perhaps you could emphasise the limited discriminatory performance of most models (usually AUC <0.7) – would such discrimination outperform clinical opinion / be useful? Those with an AUC approaching 0.9 are on small sample sizes with probably overfitting.

Response: now included in discussion section

Reviewer: 3 Johanna Damen

We would like to that the reviewer for their most insightful review of our paper. We have addressed each comment below and as a results we believe that our paper has been enhanced by your knowledge and expertise for which we thank you very much.

- What is the actual focus and aim of the review? Is it to describe the quality of the available models? Is it to teach about best practice when developing prediction models? Or is it to convince the reader that Bayesian Networks are a better method for making a prognosis? I believe these are three different papers and I suggest the authors choose one of these aims.

Response: The focus, aims and objectives of the paper are now more clearly stated i the concluding paragraph of the introduction and the title has also been amend to reflect the focus of the paper more clearly.

- From the introduction it is not clear how prediction models can be used in clinical practice. Which decisions can be made using these models for this specific disease? Please specify this.

Response: this is now clearly specified in the introduction section

- The search strategy is not fully reported. Please adjust Supplementary Material S1 to make sure the full strategy is listed, including all Boolean operators, as it is now unclear how search terms were combined. Further, there are duplicate terms in the list currently provided (e.g. ‘model’, ‘models biological’ and ‘models biologic’ are listed more than once). It is also not clear whether Mesh/Emtree terms have been used or free text. Please also add this. At this moment I do not have enough information to judge whether the search is appropriately conducted.

Response: Supplementary material S1 now states all MeSH terms and combinations of MeSH terms used in the search of each of the databases.

- Why are studies that externally validate a model without updating the model excluded? I believe these types of studies can provide most valueble information about the quality and usefulness of currently available models and therefore should not be excluded.

Response: As the focus of the paper has now been made more clear, that this was a systematic review of methodological quality of model development studies. therefore studies that externally validate a model without updating the model were excluded as they were not model development studies. This point has also been made more clearly in the methods section when inclusion and exclusion criteria are discussed.

- There are several mistakes in the flow chart (Figure 1). How is it possible that the number of records screened after duplicates removed is lower than the number of records identified in Embase? Usually this is never lower. Also the difference between the number of records identified using PubMed/Medline is very low compared to the number identified with Embase. Usually there is a difference indeed, but I have never seen a difference this large. Some of the numbers are also listed in the wrong box: the number of records screened should be 23,097 and not 263, and the number of full-text articles assessed for eligibility should be 263 and not 15. The authors can also remove the last box (studies included in quantitative synthesis) as a meta-analysis was not performed.

Response: the errors in Figure 1 have been corrected and requested changes to the diagram made.

- The conclusion is not based on the data presented in the results.

Response: the section on Bayesian modelling has been removed and instead the discussion focuses on summarising issues arising from the systematic review and the future direction of research

- There is some strange terminology used. What is a narrative systematic review? I think this is a systematic review, as it follows a systematic approach for identifying studies and collecting information. I do not see where the review is narrative. Further, with predictive we usually refer to factors or models predicting response to treatment, see for example PMID: 19383314 or the Cochrane guidance (https://methods.cochrane.org/prognosis/our-publications). To prevent confusion, it would be better to mention it as prognostic models or prediction models. Further, it is written that the authors ‘included only prognostic multivariable prediction studies where the aim was to identify a causal relationship between two or more independent variables and the outcome of prognosis (-should this be “interest”?-), to predict prognosis’. In the field of prognosis we are by definition not looking for a causal relationship. Yellow fingers can be a perfect predictor for lung cancer, but this is of course not causally related. Please remove the word causal here. Please also change univariate and multivariate to univariable and multivariable, because I believe this is what is meant by the authors.

Response: The word narrative has been removed. Prognostic models and predictive models are now referred to separately. The word causal has been removed. Univariate and multivariate have been changed to invariable and multivariable.

- A review cannot follow the PRISMA checklist, it is reported according to the PRISMA checklist (page 7, line 9)

Response: this phrasing has been corrected

- Page 10, lines 12-28, this information could be better structured below the other headings, to prevent duplicate information.

Response: the information has remained below this heading to keep with the structure of the CHARMS checklist but to facilitate flow of information and clarity of paper structure we have included a summary of the domains of the CHARMS checklist in table 1.

- Figure 2 is better readable as a bar chart. The x-axis has a strange numbering, and it now reads like in 2005, 2007 and 2008 also 1 publication was found, while I assume this was not the case (?). And why are 2017/2018 combined? I would split this up to be able to compare 2017 to the other years.

Response: Figure 2 is now a bar chart and 2017 and 2018 are presented separately.

VERSION 2 – REVIEW

REVIEWER	Dr Luke Hodgson Western Sussex Hospitals NHS FT, UK
REVIEW RETURNED	20-Mar-2019

GENERAL COMMENTS	It would be helpful for audiences to have perhaps the most commonly included variables in the abstract & briefly highlighted in the discussion. Otherwise my comments have been addressed.

REVIEWER	Johanna Damen Julius Center for Health Sciences and Primary Care, UMC Utrecht, Netherlands
REVIEW RETURNED	28-Mar-2019

GENERAL COMMENTS	I am happy to see that the authors have extensively revised their manuscript. Most of my concerns have been solved by the authors. It is now more clear what the aim of the manuscript is and it is much better readable. Now that more methodological details have been provided (e.g. the search strategy) I am better able to judge the methodological quality of the review. I have a few concerns left. Major concerns: Unfortunately I am still not convinced that the search strategy is appropriate. If I take the first line as an example ("pancreatic neoplasm" [MeSH Terms] OR Pancreatic cancer AND "prognosis"[MeSH Terms] OR prognosis[Text Word]) and type this in PubMed, then the first term ("pancreatic neoplasm" [MeSH Terms]) gives 0 hits. The correct term should be "pancreatic neoplasms" [MeSH]. Further, I believe brackets should be added: ("pancreatic neoplasm" [MeSH Terms] OR Pancreatic cancer) AND ("prognosis"[MeSH Terms] OR prognosis[Text Word]). Similar comments would apply to all other search lines. As the search is the basis for this full review, I am not convinced all relevant papers have been identified. Furthermore, the conclusion is still not based on the results and is not giving an answer to the study question. The aim of the study was to assess the methodological quality of the models, but this is not mentioned in the conclusion. Minor concern: In Supplemental Material S3, how have the ‘overall’ scores in the last two column been defined? If you follow the PROBAST guidance, many more studies would score high risk of bias.

VERSION 2 – AUTHOR RESPONSE

Reviewer Comments:

It would be helpful for audiences to have perhaps the most commonly included variables in the abstract & briefly highlighted in the discussion.

Otherwise my comments have been addressed.

Response:

Once again thank you for taking the time and effort to provide such a constructive and insightful review of our manuscript. The most commonly included variables are now included in the abstract and highlighted in the discussion.

Reviewer Comments:

I am happy to see that the authors have extensively revised their manuscript. Most of my concerns have been solved by the authors. It is now more clear what the aim of the manuscript is and it is much better readable. Now that more methodological details have been provided (e.g. the search strategy) I am better able to judge the methodological quality of the review. I have a few concerns left.

Unfortunately I am still not convinced that the search strategy is appropriate. If I take the first line as an example ("pancreatic neoplasm" [MeSH Terms] OR Pancreatic cancer AND "prognosis"[MeSH Terms] OR prognosis[Text Word]) and type this in PubMed, then the first term ("pancreatic neoplasm" [MeSH Terms]) gives 0 hits. The correct term should be "pancreatic neoplasms" [MeSH]. Further, I believe brackets should be added: ("pancreatic neoplasm" [MeSH Terms] OR Pancreatic cancer) AND ("prognosis"[MeSH Terms] OR prognosis[Text Word]). Similar comments would apply to all other search lines. As the search is the basis for this full review, I am not convinced all relevant papers have been identified.

Furthermore, the conclusion is still not based on the results and is not giving an answer to the study question. The aim of the study was to assess the methodological quality of the models, but this is not mentioned in the conclusion.

In Supplemental Material S3, how have the ‘overall’ scores in the last two column been defined? If you follow the PROBAST guidance, many more studies would score high risk of bias.

Response:

Thank you for applying your expertise to reviewing our manuscript and providing such a thoughtful and insightful review. We are very much appreciative for you help in improving our manuscript. Furthermore we would like to thank you for acknowledging the improvements made since the first version of the manuscript and we would like to acknowledge your help in achieving this.

The search strategy is now provided in the correct format as requested. The type “pancreatic neoplasm” has been corrected to “pancreatic neoplasms” and brackets have been added.

As you correctly pointed out the conclusion was not based on the results. We have changed this and the conclusion now summarizes the main methodological issues highlighted and emphasizes the need for future research to address these concerns moving forward. Once again thank you for pointing this out as we feel this has helped us to improve the manuscript.

Finally in Supplementary material 3 the overall scores were arrived at by consensus amongst all the authors following the PROBAST guidelines. As blinding was an issue across all studies we had scored these studies based on how all other aspects performed. This was noted as a footnote in the original table. As you correctly pointed out the overall assessment of bias therefore was therefore too lenient. We have therefore rescored the overall risk of bias and many more studies now have a high risk of bias, which more accurately reflects the findings of the review and more clearly conveys these results to the reader.