PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Systematic Review of Prediction Models for Delirium in the Older Adult Inpatient
AUTHORS	Lindroth, Heidi; Bratzke, Lisa; Purvis, Suzanne; Brown, Roger; Coburn, Mark; Mrkobrada, Marko; Chan, MTV; Davis, Daniel; Pandharipande, Pratik; Carlsson, Cynthia; Sanders, Robert

VERSION 1 – REVIEW

REVIEWER	Dr Annmarie Hosie University of Technology Sydney, Australia
REVIEW RETURNED	29-Sep-2017

GENERAL COMMENTS	Thank you for the opportunity to review your manuscript entitled "Systematic Review of Prediction Models for Delirium in the Older Adult Inpatient". Overall, this is a very well written and reported systematic review with a strong and clear rationale and good attention to detail and congruency between the objective/s, results and conclusions. However, there are some areas needing attention to improve the clarity of the reporting, and I have some other queries/comments, as follows: 1. The abstract has some errors of repetition – please check and revise. 2. The formatting of dates of the search, with the years provided first, is unusual – why have the dates been presented in this way? 3. I couldn't find the CHARMS checklist in Appendix A – is this referring to the Table title 'Review Protocol'? 4. What do the acronyms CDC and UN mean? (page 5). Please use full term first. 5. I am querying the justification for the decision to exclude studies conducted in palliative care populations (which was, "as these are not generalizable to an inpatient hospital setting"), as in many parts of the world, palliative care units are situated in acute hospitals. Furthermore, patients receiving palliative care are situated in almost every unit in hospitals, especially medical, geriatric and ICU settings. Lastly, of the most commonly used variables in the delirium prediction models, three (i.e. older age, impaired functional status and severe illness) are either always or commonly characteristics of patients receiving palliative care. The exclusion of people receiving palliative care from delirium studies is historically common, but doesn't make a lot of sense when delirium and dying are associated. I doubt there are many DPM studies conducted in this population, but this doesn't justify their exclusion either. Please provide a stronger justification for this particular exclusion. 6. The aim is clear in the Review Protocol, but less so in the

	body of the manuscript. Suggest it is be presented under a sub-heading 'Aim'. 7. It is hard at first to get a sense of your objectives, or the specific outcomes. Eventually I found them, but suggest it would be clearer for the reader if they were listed earlier and together. 8. The meaning of the sentence beginning' 'Two studies that included younger populations…" (page 7) is unclear. As this indicates you modified the inclusion criteria to fit these studies, it's important the rationale is clear to the reader. 9. It is right that although you included 'ICU' in some of the search terms, you found no studies conducted in the ICU setting? If so, this is worth commenting on. 10. Given that one of the aspects of the studies you examined was their statistical methods, including power calculations (and as this aspect and the limitations you found are discussed in detail later), it would be helpful to include these in Table 2. 11. The section headed 'Implications and future research' is helpful and targeted, but would benefit from further editing. I also wondered what 'AI methods and machine learning' referred to, as these ideas seemed to pop up out of nowhere. 12. At times in the body of the manuscript and especially in the abstract, an overly abbreviated writing style has been used. There are also several missing commas. Suggest a final edit by a collaborator with fresh eyes. Thanks again for this detailed and rigorous report of your work, and I wish you well with the revision and future research in this area.

REVIEWER	Diether Kramer Steiermärkische Krankenanstaltengesellschaft m.b.H., Austria
REVIEW RETURNED	09-Oct-2017

GENERAL COMMENTS	First of all this is a very detailed review on the existing research in this field. Thank you! I have only minor suggestions to (hopefully) improve the quality of your paper. 1. p4: Yow write: … expenditure of $164 billion. Where? Please add a geographical information 2. I have got the impression, that you have not adressed the problems of under- and overfitting clearly understandable. The EPV concept is a good indicator whether a model is at risk of overfitting or not, but it is only an indicator. To my knowledge the only reliable method to avoid overfitting is having a training/development and a validation and/or test data set. The validation or test data set must not be involved in the training process of the statistical model. Using a stepwise or penalised selection can help not to trap into overfitting during the training process because it reduces the complexity of the model. Anyway, whether a model is "overfitted" or not can only be recognized if the predictive ability is significantly smaller with the validation/test data. On page 18 you are writing: This effect is highlighted in the Carrasco et al.(2014) model as the AUROC decreased from the development study (0.82) [should be 0.86?] to the external validation study (0.78). I think this conclusion is not fair, because when looking at the confidence intervals it can be seen, that these models are not really different from each other. The reduced AUROC could be purely random. In addition, it should be noted that from a confidence interval range between .66 to .90 one can only draw one reliable conclusion: (too) small sample size.

	3. p11: First paragraph: All statistical methods described are logistic regression models and all of them are multivariate/multivariable. Speaking of a binomial logistic regression model means that the outcome variable is binary (e.g. 0/1). Multinomial logistic regression models are addressing multiclass problems; an ordinal regression model is a regression for an ordered dependent variable. Stepwise is just an "add-on" to select features and can help avoiding overfitting due to too many features. So 28 (unread) is the same as 26, 31, 33, 40. This means: Five studies employed a multivariate logistic model, seven additionaly apllied a stepwise feature selection and one a combinded approach. 4. p22 Implications and future research: You suggest (6) to consider AI and machine learning methods. I think you should either mention this already before or at least cite a relevant paper there (e.g. Nr 61 Newman et al)

REVIEWER	Sarah T Pendlebury Centre for Prevention of Stroke and Dementia, Nuffield Department of Clinical Neurosciences, University of Oxford and Departments of Acute Medicine and Geratology, Oxford University Hospitals NHS Trust, UK
REVIEW RETURNED	09-Oct-2017

GENERAL COMMENTS	General comments This is an interesting and useful paper reviewing existing delirium prediction models in the older adult patient. The authors have conducted a thorough review and the paper is generally easy to follow. However, in comparing the various studies, the authors should refer to the TRIPOD guidance (Moons et al, Ann Int Med 2015; 162:W1-W73) which in particular highlights the importance of case-mix/generalizability of cohorts in individual studies, method of external validation (narrow vs broad), and the issue of practicality/feasibility of scores. The discussion section regarding statistical methods to produce "better" scores needs some qualification particularly in relation to the above points. Specific comments 1. Abstract., page 2. There is repetition of the inclusion criteria which should be removed. 2. Abstract, Results, page 3. The differences in case-mix and study population have likely greater impact on measured delirium rates than the methods of identifying it (see later). 3. Abstract, conclusions, page 3. The development of "more robust models" may not be possible because of the limitations of predictive models especially in complex heterogenous patient populations and the need for such models to be pragmatic to be useful (see later). 4. Strengths and limitations, page 3. I do not think that the lack of inclusion of predictive risk factors is a limitation – this was not the focus of the review. 5. Introduction, page 4, last paragraph. It is stated that the aim was to "provide important recommendations on study design for future models". However, if this aim is to be achieved, thought needs to be given regarding generalizability and how models will be used/applied in practice (see later). These factors should therefore be assessed when reviewing available models.

	6. Methods, page 5. The terms used in the literature search seem relatively narrow. It would be usual to use MESH headings and exploded terms. This should be clarified. 7. Methods, page 6. It is not clear why the authors chose the Newcastle Ottawa criteria for assessing quality of studies (see also Results). The TRIPOD guidance was specifically developed to aid the development and also the critical appraisal of risk score studies and should be the basis of assessment of study quality. 8. Methods, statistics, page 6. There are other aspects that govern clinical utility of a model. For example, a risk score may have moderate AUC and yet may work well as a rule out test if patients with scores below a certain cut-off are highly unlikely to get the condition. Utility also depends on the consequences of missing individuals at high risk ie those not identified correctly by the score. For example, missing individuals who go on to have a stroke or cancer is arguable worse than missing those who develop delirium in whom identification of a high risk group may be helpful in targeting care. 9. Results, page 7. The methods of assessing study quality should be revised. The Newcastle Ottawa criteria would not appear correct for comparing studies of risk scores (see earlier point and also Collins and Moons; Comparing risk prediction models. BMJ 2012;344:e3186 and BMJ 2012;344:e3318). The TRIPOD guidance states the importance of the use of a representative sample and whether all consecutive participants have been included since this will determine the generalizability and interpretation of the findings. Some studies (a minority) use a consecutive prospective cohort whereas others use a selected or convenience sample. The case-mix details should therefore be included in the table of included studies and discussed in the text along with other methodological differences of relevance (see later). 10. Results, page 10. In relation to the delirium ascertainment, some studies used the managing clinician/team whereas others used external researchers who were not otherwise involved with the patients and others used nursing staff who would have been changing several times a day. Given that observation over time is helpful in diagnosing delirium, lack of continuity is likely more of an issue than the type (and frequency) of assessment and these issues should be discussed. 11. Results, page 11, top of page. The TRIPOD guidance concerning prognostic models states that there is no accepted power calculation for risk scores although it is suggested that 10 events per candidate variable would be reasonable. This therefore needs clarification here. 12. Results, page 11, variables. It would be useful to briefly list the cognitive tests used in each score (or point the reader to table 2). 13. Results, page 11, variables. The availability of variables at the point of use of the score should be examined. Some scores are derived using variables that would not be available in routine clinical practice or at the point when the score would need to be calculated to inform care.

	14. Results, page 11 and 15. List of externally validated models and predictive ability. The delirium susceptibility score should be included as an externally validated score as stated in the original publication (ref 48), (also in Table 2 and Fig 3). The score was developed from risk factors reported in the systematic review of the literature in the UK NICE guidelines and was not developed or modified/weighted on the basis of the validation cohort on which it was (externally) validated. The AUC for incident delirium was 0.81 (0.70-0.92). This score uses the same definition of cognitive impairment and of functional impairment as ref 32 (page 11). 15. Results, page 11/12. External validation. The authors should include a section on the methodology of external validation and distinguish between those studies where validation was done within the same institution ie a validation cohort is collected after the development cohort (temporal or narrow validation), and true external validation (broad or geographic validation) where a score is tested in a different institution (which is rarely done even though this is the most robust form of validation). This informs the generalizability of the score and also the interpretation of the validation AUC since this would be expected to be higher where validation occurs within cohorts collected in the same way in the same institution. Also, operationalisation of variables should be considered (see ref 32) – retaining predictive value despite modification of variable definition suggests that the variable is a powerful predictor. 16. Results. There needs to be a section added on the feasibility/applicability of scores in clinical practice. This is one of the most important factors in the clinical utility of a prognostic score (see TRIPOD). However good the AUC for a given score, it will not be used if it involves lengthy complex assessments or items not easily available in routine practice eg complex functional assessment measures. In addition, if scores are to be used routinely, they must use commonly used assessments eg the AMTS for cognition and not require the user to do a “special” non-standard test. This is particularly important if scores are to be used in non-specialist areas or at first assessment (these issues were extensively considered in ref 48). 17. Discussion, I am not convinced that the lack of multiple daily assessment is the reason for suboptimal model performance. This has not been substantiated by the data as presented ie poorer performance in models where assessment was less frequent. As stated earlier, case-mix and heterogeneity within the delirium population will be more important factors as well as overfitting of models etc. 18. Discussion, page 17. Given that the delirium susceptibility score (ref 48) is externally validated, there are therefore six scores with moderate predictive ability (see earlier points). 19. Discussion, page 17. The authors are correct to point out the issues re variable definition. However, they should note the above point regarding the importance of variable robustness to adaptation and also the need for variables to be available at the point of score use and easy to use (see earlier). As they note, it is possible to compare models when this is done within the same population (ref 32). However, the authors should justify their suggestion that

	generalisability of the results from such studies to subsequent similar populations is unclear: broad external validations of existing scores in which the model variables are operationalised/adapted and then tested in an external dataset are much more likely to be generalizable than those from narrow validations. I am not sure that keeping the same variables in a score but adapting them equates to redevelopment rather than validation. 20. Discussion, page 17. The limitations of prognostic models are not generally statistical (other than from small sample size) but more to do with case-mix issues and feasibility as well as the clinical heterogeneity of the condition being detected. In discussions re the AUC, the earlier points should be noted with respect to the robustness or otherwise of external validation therefore models that have high AUC in broad external validations (ie from cohorts from other institutions or where the model is derived using externally determined risk factors) carry more weight than similar AUC obtained using narrow validation (ie from the same institution). 21. Discussion, page 19. Re the generalizability issues, this section should be expanded in line with the above points (consecutive cohort vs selected sample, broad vs narrow external validation, use of easily obtained simple variables etc). 22. Discussion, page 19/20. The authors are correct to highlight the problems in determining whether functional impairment reflects cognitive or physical impairments or both. Of note this was examined ref 48 in which the addition of functional impairment to the model in addition to cognitive impairment did not improve the score performance. This study also showed that removal of the age criterion did not reduce score performance in line with the authors comments on the lack of additional predictive value of age. Minor points Methods, page 5. The American date format has been used which is confusing to European readers.

REVIEWER	Richard N. Jones, ScD Brown University, Providence, Rhode Island, USA
REVIEW RETURNED	14-Oct-2017

GENERAL COMMENTS	This is an excellent manuscript that is already valuable to me and my research. I can't wait to cite it. It is a wonderfully complete systematic review, for which it can serve as an exemplar. The content area and detailed and thoughtful discussion are highly relevant to delirium researchers. I have only two minor comments: 1. There does not seem to be a need to separate clinically significant from non-clinically significant prediction models based on an arbitrary AUROC of 0.75 cut off. The value of any cutoff is dubious. Whether or not a prediction model is clinically significant may be more directly informed by whether or not the prediction model included commonly clinically available predictors. This is a minor point. 2. In the discussion, penalized regression models and machine learning models are discussed as potential avenues for addressing

	methodological limitations of existing approaches, but the presentation is separate. These sections might be blended to provide a discussion of likely improvements in a single location within the document.

VERSION 1 – AUTHOR RESPONSE

Editorial Requirements:
1. Strength and limitations section (after the abstract) is presented as a list of full sentences
Response: Updated to full sentences.
Strengths and Limitations of this Study
Strengths of this systematic review include the following: the use of the PRISMA Statement and the CHARMS checklist to develop the protocol, an interprofessional authorship that provides different perspectives on delirium prediction models and a comprehensive search using multiple databases and search terms. This systematic review is limited by the focus on an older population (>60) and did not review models created in younger patients. Further, this review is limited by population focus, we did not include prediction models built in palliative care, long-term care facilities or the emergency department.

Associate Editor:
1. The only thing I felt it lacked is some discussion on the clinical usefulness. The authors call for more and better models to be developed, but not everyone thinks like that. For example, the authors of a very similar paper focusing on cardiovascular risk prediction tools say that, rather than coming up with more models, efforts should go towards externally validating and doing head to head comparisons of existing models, tailoring / combining some of these models or extending them by adding new predictors: Prediction models for cardiovascular disease risk in the general population: systematic review http://www.bmj.com/content/353/bmj.i2416
Response: The paragraph Implications and future research was updated with the following to address the potential creation of a meta-model with the currently identified delirium prediction models.
Pg24: “Two avenues may be pursued for future studies. The first avenue involves model aggregation; currently available delirium prediction models would be combined into a meta-model through stacked regression in a new cohort of participants. This method would update currently published models to a new population, furthering generalizability and bolstering broad external validation.1 Variable definition could be harmonized in the meta-model with the intention to use variables that are readily available and feasible for routine practice. This method would further delirium prediction for those with dementia-level pre-existing cognitive impairment as well as examine the individual contributions of functional impairment due to physical conditions, cognitive impairment or age through model re-fitting. Nonetheless, a future meta-model would continue presently identified limitations such as exclusion of the spectrum of cognition. The second avenue should focus on the development and broad validation of delirium prediction models exploring the use of simple cognitive tests that would be inclusive to mild cognitive impairment and sensitive to the spectrum of cognition. Further, future models should consider development of dynamic predictive models using advanced statistical methods such as Bayesian Networks, artificial intelligence, and machine learning as these methods have shown to improve models built using standard logistic regression.2 3”
Reviewer 1: Dr. Annmarie Hosie
Thank you for the opportunity to review your manuscript entitled "Systematic Review of Prediction Models for Delirium in the Older Adult Inpatient".

Overall, this is a very well written and reported systematic review with a strong and clear rationale and good attention to detail and congruency between the objective/s, results and conclusions.
Response: Thank you for your compliments on the overall manuscript. Your comments were very helpful in

improving this manuscript.

However, there are some areas needing attention to improve the clarity of the reporting, and I have some other queries/comments, as follows:

1. The abstract has some errors of repetition – please check and revise.

Response: Removed abstract section titled “eligibility”, removed repetition of inclusion/exclusion criteria.

2. The formatting of dates of the search, with the years provided first, is unusual-why have the dates been presented in this way?

Response: We used the international standard date notation, this is the style used by the BMJ. It has been updated to 1st January 1990 to 31st December 2016.

3. I couldn’t find the CHARMS checklist in Appendix A is this referring to the Table title “Review Protocol”?

Response: We mistakenly did not upload the CHARMS checklist, this is now part of Appendix A.

4. What do the acronyms CDC and UN mean? (page 5). Please use full terms first.

Response: This has been corrected to read the following:

Pg 5: “This review included studies focused on 1) older adult (> 60 years) population, (the U.S. Center for Disease Control and Prevention and United Nations define an older adult as 60 years of age and older).”

5. I am querying the justification for the decision to exclude studies conducted in palliative care populations (which was, “as these are not generalizable to an inpatient hospital setting”), as in many parts of the world, palliative care units are situated in acute hospitals. Furthermore, patients receiving palliative care are situated in almost every unit in hospitals, especially medical, geriatric and ICU settings. Lastly, of the most commonly used variables in the delirium prediction models, three (i.e. older age, impaired functional status and severe illness) are either always or commonly characteristics of patients receiving palliative care. The exclusion of people receiving palliative care from delirium studies is historically common, but doesn’t make a lot of sense when delirium and dying are associated. I doubt there are many DPM studies conducted in this population, but this doesn’t justify their exclusion either. Please provide a stronger justification for this particular exclusion.

Response: As palliative care delirium has unique challenges, we felt that it would be best addressed with a specific focus in an alternate review. While we agree with your point that patients receiving palliative care are situated in almost every unit in hospital, most palliative care patients are likely admitted as medical or surgical patients and as their prognosis evolves, their status may change to palliative. Further, therapeutic interventions may be different for delirium in palliative care patients as demonstrated by the recent stud published by Hui et al. (2017) reporting that use of a benzodiazepine reduced agitation and increased perceived comfort. This is in opposition to the Pain, Agitation and Delirium guidelines. In the above scenario, this systematic review is inclusive of these patients. We have updated the exclusion language to state the following:

Methods

Pg5: ..studied a different patient population (i.e. emergency department, skilled nursing facilities, palliative care, and hospice) as these are unique patient populations with characteristics requiring specific foci and are not readily generalizable to a medical or surgical inpatient hospital setting. Further, recommended therapies for treatment of delirium symptoms vary between the populations4 5.

Further, we have added this narrowed population focus to our limitations.

Strengths and Weaknesses of this study

Pg24: “… Further, this review is limited by population focus. We did not include prediction models built in palliative care, long-term care facilities, or the emergency department.”

6. The aim is clear in the Review Protocol, but less so in the body of the manuscript. Suggest it is be presented under a sub-heading 'Aim'.
Response: We have inserted the subheading "Aim" on page 4.
7. It is hard at first to get a sense of your objectives, or the specific outcomes. Eventually I found them, but suggest it would be clearer for the reader if they were listed earlier and together.
Response: We have inserted the subheading "Outcomes" on page 5.
8. The meaning of the sentence beginning' 'Two studies that included younger populations…" (page 7) is unclear. As this indicates you modified the inclusion criteria to fit these studies, it's important the rationale is clear to the reader.
Response: We further clarified this sentence by the following language:
Results
Pg 7: The inclusion criteria were modified for two studies that developed models in younger populations, but these models were externally validated in the target population of this review (age > 60).

9. It is right that although you included 'ICU' in some of the search terms, you found no studies conducted in the ICU setting? If so, this is worth commenting on.
Response: We did not identify any studies that focused on adults >60yo and were conducted in the ICU. We added the following to clarify:
Results
Pg 7: None of the identified studies focused on critical care patients.

10. Given that one of the aspects of the studies you examined was their statistical methods, including power calculations (and as this aspect and the limitations you found are discussed in detail later), it would be helpful to include these in Table 2.
Response: This information was added to Table 1, in the 2nd column. We updated Table 1 instead of Table 2 because one of the development studies reported their power analysis and Table 2 is solely focused on the externally validated models.
11. The section headed 'Implications and future research' is helpful and targeted, but would benefit from further editing. I also wondered what 'AI methods and machine learning' referred to, as these ideas seemed to pop up out of nowhere.
Response 1: Implications and future research has been edited and clarified further. It now reads as follows:
Discussion
Implications and future research
Pg24: "Two avenues may be pursued for future studies. The first avenue involves model aggregation; currently available delirium prediction models would be combined into a meta-model through stacked regression in a new cohort of participants. This method would update currently published models to a new population, furthering generalizability and bolstering broad external validation.1 Variable definition could be harmonized in the meta-model with the intention to use variables that are readily available and feasible for routine practice. This method would further delirium prediction for those with dementia-level pre-existing cognitive impairment as well as examine the individual contributions of functional impairment due to physical conditions, cognitive impairment or age through model re-fitting. Nonetheless, a future meta-model would continue presently identified limitations such as exclusion of the spectrum of cognition. The second avenue should focus on the development and broad validation of delirium prediction models exploring the use of simple cognitive tests that would be inclusive to mild cognitive impairment and sensitive to the spectrum of cognition. Further, future models should consider development of dynamic predictive models using advanced statistical methods such as Bayesian Networks, artificial intelligence, and machine learning as these methods have shown to improve models built using standard logistic regression.2 3
We suggest the following broad principles for use in future studies: (1) Delirium prediction models should be developed only using data available prior to the onset of delirium and likely should be

focused in specific populations depending on whether the precipitating event has occurred or not; (2) should include structured, twice daily assessment (regardless of weekends) using validated tools and trained research staff to identify delirium; (3) include variables and assessments that are readily available in clinical practice and are feasible to administer without extensive training or interpretation where possible and not to exclude a more informative variable; (4) model development and validation should follow rigorous methods outlined by Steyerberg (2009)6 and Steyerberg and Vergouwe (2014)7 including strategies to counter low sample size and overly optimistic model performance, the use of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to assess model fit, and consider broad validations to expand case-mix and generalizability; and (5) adhere to strict guidelines as outlined by The TRIPOD Statement for statistical performance reporting including calibration and clinical utility statistics.6-11"
Response 2: The authors appreciate this point and the concept of AI methods and machine learning has been introduced earlier in the discussion section, on page 20, the new text is below:
Discussion
Pg21"Further, future studies may benefit from the incorporation of advanced statistical techniques such as Bayesian Networks and machine learning that have shown to improve the performance of previous prediction models that were built using standard logistic regression.2 12 These methods facilitate the exploration of complex interactions between risk factors as well as adapt to changing patient conditions, allowing for dynamic models."

12. At times in the body of the manuscript and especially in the abstract, an overly abbreviated writing style has been used. There are also several missing commas. Suggest a final edit by a collaborator with fresh eyes.
Response: This was addressed.
Reviewer #2: Diether Kramer, Austria
First of all this is a very detailed review on the existing research in this field. Thank you!
I have only minor suggestions to (hopefully) improve the quality of your paper.
Response: Thank you for your review and helpful suggestions which do improve the quality of our paper.
1. p4: Yow write: … expenditure of $164 billion. Where? Please add a geographical information
Response: Geographical information was added.
Pg4: Delirium has been independently associated with increased mortality, morbidity in terms of impaired cognition and functional disability along with an estimated annual U.S. expenditure of $152 billion.

2. I have got the impression, that you have not addressed the problems of under-and overfitting clearly understandable. The EPV concept is a good indicator whether a model is at risk of overfitting or not, but it is only an indicator. To my knowledge the only reliable method to avoid overfitting is having a training/development and a validation and/or test data set. The validation or test data set must not be involved in the training process of the statistical model. Using a stepwise or penalised selection can help not to trap into overfitting during the training process because it reduces the complexity of the model. Anyway, whether a model is "overfitted" or not can only be recognized if the predictive ability is significantly smaller with the validation/test data. On page 18 you are writing: This effect is highlighted in the Carrasco et al.(2014) model as the AUROC decreased from the development study (0.82) [should be 0.86?] to the external validation study (0.78). I think this conclusion is not fair, because when looking at the confidence intervals it can be seen, that these models are not really different from each other. The reduced AUROC could be purely random. In addition, it should be noted that from a confidence interval range between .66 to .90 one can only draw one reliable conclusion: (too) small sample size
Response: We appreciate this comment and have further clarified the instability introduced due to insufficient EPV on page 18 in the results section and on page 20 in the discussion section. This text is pasted below. We have removed the sentence "This effect is highlighted in the Carrasco et

al.(2014) model as the AUROC decreased from the development study (0.82) [should be 0.86?] to the external validation study (0.78).” as we are in agreement that this likely occurred due to a small sample size. As re-iterated by the CHARMS checklist and experts on predictive modeling (Moons and Steyerberg) we feel it is important to include the reporting and discussion of EPV so future studies can use the recommended strategies to counteract low EPV.

Results

Pg18:” Events per variable (EPV) were examined in each of the fourteen externally validated models. Models estimating more parameters than events in a 1:10 ratio are at risk of statistical overfitting, potentially leading to overly optimistic model performance.6 7 13-15 In 14 models with external validation, four had fewer than optimum events for the number of parameters estimated in the development stage of the models.16-19 Five had fewer than optimum events in the external validation stage.18-22 Two models did not reach optimum events for the number of parameters in either the development or the external validation studies.18 19 Various statistical techniques such as shrinkage procedures, the use of lasso or penalized regression and internal validation methods are suggested to counter the effects of low EPV.13 23 24 Techniques such as statistical shrinkage procedures and internal validation are recommended to counter the effects of a low sample size and insufficient EPV. None of the identified studies report use of statistical shrinkage procedures. Five studies applied internal validation techniques in the development stage of their model to establish stability within their model.16 25-28”

Discussion

“Pg20: Model underperformance may be explained by low powered studies, insufficient events per variable (EPV) as well as the use of univariate analyses and stepwise regression to select predictive variables for inclusion into models. Although these are common methods to use for model development and may counter the effects of insufficient EPV, each approach has significant drawbacks.23 Univariate analysis may reduce predictive ability by inclusion of variables that are not independent of each other, and stepwise regression disadvantages include conflation of p-values and a biased estimation of coefficients.6 9 24 29 While EPV was originally adapted to ensure stability in regression covariates, it has been identified as an important component to predictive model stability and reproducibility due to the result of overfitting.9 24 30 Ogundimu et al. (2016) demonstrate this effect by simulating models with EPV of 2, 5, 10, 15, 20, 25 and 50. Stability of models increased as the EPV increased and models including predictors with low population prevalence required >20 EPV.31 The degree of model overfitting should be assessed through calibration statistics and forms of internal validation such as bootstrapping. Future studies should consider the use of statistical methods to counter low EPV including the application of statistical shrinkage techniques and penalised regression using ridge or lasso regression.6 7 14 24 32 Further, future studies may benefit from the incorporation of advanced statistical techniques such as Bayesian Networks and machine learning that have shown to improve the performance of previous prediction models that were built using standard logistic regression.12 14 These methods facilitate the exploration of complex interactions between risk factors as well as adapt to changing patient conditions, allowing for a dynamic model.”

3. p11:First paragraph: All statistical methods described are logistic regression models and all of them are multivariate/multivariable. Speaking of a binomial logistic regression model means that the outcome variable is binary (e.g. 0/1). Multinomial logistic regression models are addressing multiclass problems; an ordinal regression model is a regression for an ordered dependent variable. Stepwise is just an “add-on” to select features and can help avoiding overfitting due to too many features. So 28 (unread) is the same as 26, 31, 33, 40. This means: Five studies employed a multivariate logistic model, seven additional applied a stepwise feature selection and one a combined approach.

Response: This has been updated per your suggestion.

Results

Pg 12: Sixteen studies employed a form of logistic regression. Twelve of these models applied a stepwise regression approach.16 18 19 21 26 28 33-38 Three applied a stepwise forward selection

process,16 18 21 two employed a stepwise backward selection process28 38 and one used a combined approach.19 Statistical methods used for model building are further outlined in Table 1.

4. p22 Implications and future research: You suggest (6) to consider AI and machine learning methods. I think you should either mention this already before or at least cite a relevant paper there (e.g. Nr 61 Newman et al).
Response: Per your suggestion, AI and machine learning are mentioned earlier in the discussion section, on page 21 and relevant papers are cited both on page 21 and 24.
Reviewer #3, Dr. Sarah Pendlebury
1. This is an interesting and useful paper reviewing existing delirium prediction models in the older adult patient. The authors have conducted a thorough review and the paper is generally easy to follow. However, in comparing the various studies, the authors should refer to the TRIPOD guidance (Moons et al, Ann Int Med 2015; 162:W1-W73) which in particular highlights the importance of case-mix/generalizability of cohorts in individual studies, method of external validation (narrow vs broad), and the issue of practicality/feasibility of scores. The discussion section regarding statistical methods to produce "better" scores needs some qualification particularly in relation to the above points.
Response: The authors appreciate this recommendation and thank the reviewer for the thorough comments. We have referenced the TRIPOD8 9 guidance statement to address methods of external validation, case-mix/generalizability of cohorts in individual studies and practicality/feasibility of scores. Since the TRIPOD guidelines focus on the reporting of research and do not "prescribe" to know how to develop or validate a prediction model. (pg59 statement), we have also referenced the CHARMS checklist to guide this review.
2. Abstract., page 2. There is repetition of the inclusion criteria which should be removed.
Response: Corrected.
3. Abstract, Results, page 3. The differences in case-mix and study population have likely greater impact on measured delirium rates than the methods of identifying it (see later).
Response: Dr. Pendlebury is correct that case-mix is important in predisposition to delirium, but we would argue that diagnosis is key. This comment spurred us to analyze the reported delirium rates per study population. This analysis did not demonstrate a statistical difference in delirium rates between medical, medical/surgical and surgical populations. Nonetheless we have included the importance of case-mix in the discussion section as shown below as we agree that this is likely an important factor.

Discussion
"Pg 20: While case-mix between populations may impact observed delirium rates, we believe it would be advantageous for future studies to incorporate systematic, frequent and consistent delirium assessments."

4. Abstract, conclusions, page 3. The development of "more robust models" may not be possible because of the limitations of predictive models especially in complex heterogenous patient populations and the need for such models to be pragmatic to be useful (see later).
Response: While the authors agree that predictive models could be limited due to complex heterogenous patient populations, future models may improve with the outlined recommendations in this review. Models should first be predictive then be pragmatic.
5. Strengths and limitations, page 3. I do not think that the lack of inclusion of predictive risk factors is a limitation – this was not the focus of the review.
Response: We have removed this as a limitation.
6. Introduction, page 4, last paragraph. It is stated that the aim was to "provide important recommendations on study design for future models". However, if this aim is to be achieved, thought needs to be given regarding generalizability and how models will be used/applied in practice (see later). These factors should therefore be assessed when reviewing available models.

Response: We appreciate this comment and have interweaved narrow versus broad validation, as defined by TRIPOD guidelines, as well as clinical utility, throughout the manuscript. We have also included recommendations for future studies to report on clinical utility and how models can be incorporated into practice.

Clinical Utility

Results

"Pg18: Clinical utility of a prediction model may be evaluated through several different statistical metrics including odds ratios, relative risk, sensitivity and specificity, receiver operator curves, R squared and integrated discrimination improvement indices as well as the clinical utility curve statistic.11 Six externally validated delirium prediction model studies reported odds ratios or relative risk statistics evaluating the highest risk stratification point.18-20 28 39 40 Seven studies reported sensitivity and specificity21 37 38 40-43 and one study reported the rate of true positives and false positives.44 None of the identified studies reported decision curse analysis or clinical utility curve analysis. While the majority of studies selected variables that were either routinely used in practice or were feasible to administer, two studies developed delirium prediction models based on data routinely entered into the electronic health record to increase feasibility of use.25 44 Pendlebury et al. (2016) adapted variable definition and use to match routine clinical assessment while externally validating four delirium prediction models.43 Moerman et al. reported feasibility and reliability statistics following the incorporation of the risk prediction tool into practice.41"

Clinical utility is further mentioned in the discussion section regarding feasible cognitive tests (pg. 21)

Discussion

Pg23: "Further exploration into isolated cognitive tests that are feasible to administer in a clinical setting as well as sensitive to the spectrum of cognitive impairment may enhance delirium prediction."

And is discussed in a stand-alone paragraph in the discussion section:

Discussion

"Pg23 The clinical utility of a prediction model is dependent on both its efficacy at predicting those at risk and feasibility hence both must be considered when building and validating a model. Clinical utility is compromised by efficacious models that are not feasible. Conversely, a feasible model that is not effective at identifying those at risk also lacks clinical utility. To this end, model derivation must focus on building an effective model. The next aspect that must be considered is the ability to enhance clinical care. Predicting individuals at high risk is clearly important, but to an experienced clinician, delirium may already be anticipated. Maximum value may be obtained by aiding in prediction of moderate risk patients, where the risk of delirium may be more ambiguous."

And on pg.25

Discussion

…"adhere to strict guidelines as outlined by The TRIPOD Statement for statistical performance reporting including calibration and clinical utility…"

7. Methods, page 5. The terms used in the literature search seem relatively narrow. It would be usual to use MESH headings and exploded terms. This should be clarified.

Response: MESH terms were referenced during the creation of the keyword terms for this systematic review. MESH terms for delirium were not added to the PubMed database until 2017. At the time of list development, in 2015, the available MESH terms would not have significantly contributed to the already large and broad keyword list.

8. Methods, page 6. It is not clear why the authors chose the Newcastle Ottawa criteria for assessing quality of studies (see also Results). The TRIPOD guidance was specifically developed to aid the development and also the critical appraisal of risk score studies and should be the basis of assessment of study quality.

Response: The Newcastle Ottawa scale was chosen as most included studies are prospective cohort studies. Per your suggestion, we did contact the TRIPOD authors for guidance and as a result applied the CHARMS checklist for risk of bias. The text is updated on pg 6 & 7 with the following text. A figure was added as well.

Pg6: Risk of bias was assessed through the CHARMS checklist.24

Pg7: Risk of bias was assessed using the CHARMS checklist24 and results are shown in Figure 2.

9. Methods, statistics, page 6. There are other aspects that govern clinical utility of a model. For example, a risk score may have moderate AUC and yet may work well as a rule out test if patients with scores below a certain cut-off are highly unlikely to get the condition. Utility also depends on the consequences of missing individuals at high risk ie those not identified correctly by the score. For example, missing individuals who go on to have a stroke or cancer is arguable worse than missing those who develop delirium in whom identification of a high risk group may be helpful in targeting care.

Response: We agree that clinical utility is difficult to address largely due to the unavailability of objective evaluation tools and the dearth of delirium prediction models currently incorporated into practice and/or the lack of published data showing their incorporation into practice. From our perspective, the main value of a prediction score would be enhanced decision making in those patients were delirium occurrence is not readily anticipated. As an example, an adult with a large traumatic accident requiring emergent surgery and is admitted to the ICU on mechanical ventilation has a high likelihood of becoming delirious versus an older adult approximately 65-75 years of age with deficits in executive function and an underlying infectious process. This older adult would be missed by current models because he/she is not cognitively impaired to the level of dementia and is not old enough as scored by existing delirium prediction models (>80). While we agree that missing a stroke or cancer is not desirable, we are focused on delirium prevention. In order to prevent delirium in those that may be at risk, but do not have the well-established risk factors, such as dementia, it is important to stratify individuals appropriately.

To address clinical utility, we have interweaved clinical utility throughout the review and new language is written, please refer to our previous response, found on pg8, comment #6. Clinical utility metrics have been added to Table 2.

10. Results, page 7. The methods of assessing study quality should be revised. The Newcastle Ottawa criteria would not appear correct for comparing studies of risk scores (see earlier point and also Collins and Moons; Comparing risk prediction models. BMJ 2012;344:e3186 and BMJ 2012;344:e3318). The TRIPOD guidance states the importance of the use of a representative sample and whether all consecutive participants have been included since this will determine the generalizability and interpretation of the findings. Some studies (a minority) use a consecutive prospective cohort whereas others use a selected or convenience sample. The case-mix details should therefore be included in the table of included studies and discussed in the text along with other methodological differences of relevance (see later).

Response: The CHARMS checklist for Risk of Bias is incorporated into this review, reference Figure 2. Sampling method was recorded and is shown in Table 1. Text is also inserted on page 7.

Results

Pg 7: “Nineteen studies used consecutive sampling methods,16-22 27 33 34 36 39-46 two of these were part of a randomized control trial.39 41

11. Results, page 10. In relation to the delirium ascertainment, some studies used the managing clinician/team whereas others used external researchers who were not otherwise involved with the patients and others used nursing staff who would have been changing several times a day. Given that observation over time is helpful in diagnosing delirium, lack of continuity is likely more of an issue than the type (and frequency) of assessment and these issues should be discussed.

Response: We agree with the reviewer that observation over time is helpful as well as the employment of professionally trained, consistent study staff to assess for delirium in research. From our perspective the frequency of assessment contributes and bolsters the observations made over time and contributes to continuity due to the acute and fluctuating nature of delirium. We have addressed the concerns of consistency in the discussion section on page 19 (below).

Discussion
Pg 19-20: “Lastly, assessment of the outcome variable, delirium, was largely non-systematic, once daily, and avoided weekends. In the studies that assessed delirium more than once per day, the assessment was performed by routine clinical staff, decreasing consistency. This is a major limitation for an acute condition that fluctuates, may occur suddenly and is dependent on precise, objective assessment. While case-mix between populations may impact observed delirium rates, we believe it would be advantageous for future studies to incorporate systematic, frequent and consistent delirium assessments.”

12. Results, page 11, top of page. The TRIPOD guidance concerning prognostic models states that there is no accepted power calculation for risk scores although it is suggested that 10 events per candidate variable would be reasonable. This therefore needs clarification here.

Response: The TRIPOD Statement was developed as a recommendation and guide for reporting of research focused on developing or validating prediction models (pg 59 of TRIPOD Statement). While TRIPOD does state that 10 events per variable would be reasonable, the authors of TRIPOD state do not prescribe on how to develop or validate a prediction model. The authors of this review have referred to the CHARMS checklist for guidance on events per variable as it is designed to assist authors in the critical appraisal of prediction modeling studies. CHARMS details and further substantiates event per variable requirements by outlining sample size considerations and effects of insufficient events per variable on model development and validation.

13. Results, page 11, variables. It would be useful to briefly list the cognitive tests used in each score (or point the reader to table 2).

Response: Reader has been pointed to Table 2. This text is inserted on page 12.

14. Results, page 11, variables. The availability of variables at the point of use of the score should be examined. Some scores are derived using variables that would not be available in routine clinical practice or at the point when the score would need to be calculated to inform care.

Reponse: We discuss the limitations of available data and issues with defining them in detail before delirium diagnosis in discussion section, on page 20. Further, this was addressed in the CHARMS Risk of Bias assessment, Figure 2.

15. Results, page 11 and 15. List of externally validated models and predictive ability. The delirium susceptibility score should be included as an externally validated score as stated in the original publication (ref 48), (also in Table 2 and Fig 3). The score was developed from risk factors reported in the systematic review of the literature in the UK NICE guidelines and was not developed or modified/weighted on the basis of the validation cohort on which it was (externally) validated. The AUC for incident delirium was 0.81 (0.70-0.92). This score uses the same definition of cognitive impairment and of functional impairment as ref 32 (page 11).

Response: The referred to study has been added as the fourteenth externally validated delirium prediction model. It is added to Table 2 and Figure 4 (was figure 3).

16. Results, page 11/12. External validation. The authors should include a section on the methodology of external validation and distinguish between those studies where validation was done within the same institution ie a validation cohort is collected after the development cohort (temporal or narrow validation), and true external validation (broad or geographic validation) where a score is tested in a different institution (which is rarely done even though this is the most robust form of validation). This informs the generalizability of the score and also the interpretation of the validation AUC since this would be expected to be higher where validation occurs within cohorts collected in the same way in the same institution. Also, operationalisation of variables should be considered (see ref 32) – retaining predictive value despite modification of variable definition suggests that the variable is a powerful predictor.

Response: A section on type of external validation has been added to the results section on page 11 and the text is below. This information is also added into Table 2.

Results
Pg 11: “Per TRIPOD reporting guidelines, validation studies were categorized into type; narrow validation refers to the same investigators subsequently collecting an additional patient cohort,

following the development cohort, and broad validation refers to a validation cohort sampled from a different hospital or country.8-10 As interpretation of validation studies is dependent on case-mix,47 it is important to note that eight of the fourteen externally validated models are categorized as narrow validations.18-21 28 38 41 42 Further information is outlined in Table 2."

17. Results. There needs to be a section added on the feasibility/applicability of scores in clinical practice. This is one of the most important factors in the clinical utility of a prognostic score (see TRIPOD). However good the AUC for a given score, it will not be used if it involves lengthy complex assessments or items not easily available in routine practice eg complex functional assessment measures. In addition, if scores are to be used routinely, they must use commonly used assessments eg the AMTS for cognition and not require the user to do a "special" non-standard test. This is particularly important if scores are to be used in non-specialist areas or at first assessment (these issues were extensively considered in ref 48).

Response: A section on clinical utility has been added to the results section and discussion section as discussed and outlined previously. We agree that the feasibility/applicability of scores in clinical practice is one of the most important factors. However, as mentioned previously, we are limited by the lack of objective assessment measures to apply in evaluating whether a model is clinical feasible or applicable. We do state in the discussion the importance of using variables that are feasible "pg 22…simple cognitive tests as employed by Fong et al. (2015), as a variable may increase the detection and prevalence of cognitive impairment as a variable thus increasing its predictive power. Further exploration into isolated cognitive tests that are feasible to administer in a clinical setting as well as sensitive to the spectrum of cognitive impairment may enhance delirium prediction." And have added the following recommendation in the "Implications and Future Research" Section:

Discussion:

Pg:25"(3) should consider inclusion of variables and assessments that are readily available in clinical practice and are feasible to administer without extensive training or interpretation,"

18. Discussion, I am not convinced that the lack of multiple daily assessment is the reason for suboptimal model performance This has not been substantiated by the data as presented ie poorer performance in models where assessment was less frequent. As stated earlier, case-mix and heterogeneity within the delirium population will be more important factors as well as overfitting of models etc.

Response: We disagree with the reviewer on this point. We believe that the lack of multiple daily assessments does impact model performance. As one of the core features of delirium is an acute presentation with a fluctuating nature, frequent assessment is important. This comment spurred us to investigate as a proof of concept whether model performance, measured by AUROC, was different dependent on the frequency of delirium assessment. We found a significant statistical difference in model performance (anova, p=0.006), between those models that assessed delirium 2xday and those that assessed delirium 1xday (post-hoc, p=0.01). Studies were included if they were externally validated and published their AUROC.

We did reorder our first paragraph of the discussion section. Our perspective is that the frequency of delirium assessment does influence model performance, as shown in our analysis, and it would be best practice moving forward for future studies to adopt an approach of assessing delirium twice daily. However, to respect this reviewers concern, we have moved the limitation of delirium assessment to the third addressed limitation, and moved overall study design and reporting as our first, and main, limitation.

DISCUSSION

"Pg19: This review identified moderate predictive ability (AUROC 0.52-0.94) in fourteen externally validated delirium prediction models with eight out of fourteen models using narrow validation. However, three main limitations were identified. First, study design, application, and reporting of statistical methods appear inadequate. Data collection overlapped with the initial diagnosis of delirium

in the highest performing model as well as in two other included studies, likely exaggerating model performance.24 38 42 48 Low EPV combined with limited application of internal validation techniques contributed to an increased risk of bias and likely the creation of overly optimistic models.8-10 24 Second, broad variable definitions, particularly in functional and cognitive abilities, may have led to overlapping data capture. For example, Pendlebury et al. (2016) demonstrated this possible effect in the development of the Susceptibility Score, model performance did not improve with the addition of functional impairment to a model that already included cognitive impairment and age.40 Lastly, assessment of the outcome variable, delirium, was largely non-systematic, once daily, and avoided weekends. In the studies that assessed delirium more than once per day, the assessment was performed by routine clinical staff, decreasing consistency. This is a major limitation for an acute condition that fluctuates, may occur suddenly and is dependent on precise, objective assessment. While case-mix between populations may impact observed delirium rates, we believe it would be advantageous for future studies to incorporate systematic, frequent and consistent delirium assessments."

19. Discussion, page 17. Given that the delirium susceptibility score (ref 48) is externally validated, there are therefore six scores with moderate predictive ability (see earlier points).

Response: This has been corrected and ref 48 is included as externally validated.

20. Discussion, page 17. The authors are correct to point out the issues re variable definition. However, they should note the above point regarding the importance of variable robustness to adaptation and also the need for variables to be available at the point of score use and easy to use (see earlier). As they note, it is possible to compare models when this is done within the same population (ref 32). However, the authors should justify their suggestion that generalisability of the results from such studies to subsequent similar populations is unclear: broad external validations of existing scores in which the model variables are operationalised/adapted and then tested in an external dataset are much more likely to be generalizable than those from narrow validations. I am not sure that keeping the same variables in a score but adapting them equates to redevelopment rather than validation.

Response: The variable definition point was refined further to highlight the potential for overlapping data collection and the mention of redevelopment rather than validation was removed since it was no longer a pertinent point. Broad validation studies were included as a recommendation for future research.

21. Discussion, page 17. The limitations of prognostic models are not generally statistical (other than from small sample size) but more to do with case-mix issues and feasibility as well as the clinical heterogeneity of the condition being detected. In discussions re the AUC, the earlier points should be noted with respect to the robustness or otherwise of external validation therefore models that have high AUC in broad external validations (ie from cohorts from other institutions or where the model is derived using externally determined risk factors) carry more weight than similar AUC obtained using narrow validation (ie from the same institution).

Response: The authors agree with the reviewer that the AUC is not the sole metric for model evaluation and narrow versus broad validation was incorporated into the methods and discussion sections. Nonetheless, calibration statistics are needed to fully evaluate the fit of the model in the new validation population. TRIPOD and CHARMS both report the importance of this metric. Of the seven models that did broad evaluation, only two of these studies reported calibration statistics. We do not feel we can give more weight to AUROC resulting from broad validations without calibration metrics to further substantiate their fit.

22. Discussion, page 19. Re the generalizability issues, this section should be expanded in line with the above points (consecutive cohort vs selected sample, broad vs narrow external validation, use of easily obtained simple variables etc).

Response: We have incorporated the importance of narrow vs broad, clinical utility, easily obtained simple variables throughout the results and discussion section. On page 23, these concepts have been highlighted. Please refer to this document: Pg3-4, item number 11 – under Reviewer #1 to read the updated paragraphs.

23. Discussion, page 19/20. The authors are correct to highlight the problems in determining whether functional impairment reflects cognitive or physical impairments or both. Of note this was examined ref 48 in which the addition of functional impairment to the model in addition to cognitive impairment did not improve the score performance. This study also showed that removal of the age criterion did not reduce score performance in line with the authors comments on the lack of additional predictive value of age.

Response: This is a great point and has been incorporated and referenced in the discussion section.

Discussion:

Pg19: “For example, Pendlebury et al. (2016) demonstrated this possible effect in the development of the Susceptibility Score, model performance did not improve with the addition of functional impairment to a model that already included cognitive impairment and age.40”

Pg22: “This effect was demonstrated by Pendlebury et al. (2016), an improved AUROC resulted when age was removed from the prediction model (0.81 to 0.84).40”

24. Methods, page 5. The American date format has been used which is confusing to European readers.

Response: We have applied the BMJ, international date format. We changed to 1 January 1990 and 31 December 2016.

Reviewer #4, Dr. Richard Jones

This is an excellent manuscript that is already valuable to me and my research. I can't wait to cite it. It is a wonderfully complete systematic review, for which it can serve as an exemplar. The content area and detailed and thoughtful discussion are highly relevant to delirium researchers.

Response: Thank you for your compliments and enthusiasm for our systematic review. We hope this can serve as an exemplar and inform future delirium research.

1. There does not seem to be a need to separate clinically significant from non-clinically significant prediction models based on an arbitrary AUROC of 0.75 cut off. The value of any cutoff is dubious. Whether or not a prediction model is clinically significant may be more directly informed by whether or not the prediction model included commonly clinically available predictors. This is a minor point.

Response: The authors appreciate this point and the arbitrary cut-off has been removed.

2. In the discussion, penalized regression models and machine learning models are discussed as potential avenues for addressing methodological limitations of existing approaches, but the presentation is separate. These sections might be blended to provide a discussion of likely improvements in a single location within the document.

Response: The implications and future research has been re-structured to include a more congruent discussion on potential areas for improvement. Please refer to this document: Pg3-4, item number 11 – under Reviewer #1 to read the updated paragraphs.

References:

1. Debray TP, Koffijberg H, Nieboer D, et al. Meta-analysis and aggregation of multiple published prediction models. Stat Med 2014;33(14):2341-62. doi: 10.1002/sim.6080 [published Online First: 2014/04/23]

2. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Plos One 2017;12(4):e0174944. doi: 10.1371/journal.pone.0174944 [published Online First: 2017/04/05]

3. Kim SY, Moon SK, Jung DC, et al. Pre-operative prediction of advanced prostatic cancer using clinical decision support systems: accuracy comparison between support vector machine and artificial neural network. Korean journal of radiology 2011;12(5):588-94. doi: 10.3348/kjr.2011.12.5.588 [published Online First: 2011/09/20]

4. Barr J, Fraser GL, Puntillo K, et al. Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the intensive care unit. Crit Care Med 2013;41(1):263-306. doi: 10.1097/CCM.0b013e3182783b72 [published Online First: 2012/12/28]
5. Hui D, Frisbee-Hume S, Wilson A, et al. Effect of Lorazepam With Haloperidol vs Haloperidol Alone on Agitated Delirium in Patients With Advanced Cancer Receiving Palliative Care: A Randomized Clinical Trial. Jama 2017;318(11):1047-56. doi: 10.1001/jama.2017.11468 [published Online First: 2017/10/05]
6. Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating: New York : Springer, [2009] ©2009 2009.
7. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European heart journal 2014;35(29):1925-31. doi: 10.1093/eurheartj/ehu207 [published Online First: 2014/06/06]
8. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine 2015;162(1):W1-73. doi: 10.7326/m14-0698 [published Online First: 2015/01/07]
9. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ (Clinical research ed) 2015;350:g7594. doi: 10.1136/bmj.g7594 [published Online First: 2015/01/09]
10. Moons KG, Altman DG, Reitsma JB, et al. New Guideline for the Reporting of Studies Developing, Validating, or Updating a Multivariable Clinical Prediction Model: The TRIPOD Statement. Advances in anatomic pathology 2015;22(5):303-5. doi: 10.1097/pap.0000000000000072 [published Online First: 2015/08/12]
11. Campbell DJ. The clinical utility curve: a proposal to improve the translation of information provided by prediction models to clinicians. BMC research notes 2016;9:219. doi: 10.1186/s13104-016-2028-0 [published Online First: 2016/04/16]
12. Strobl AN, Vickers AJ, Van Calster B, et al. Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators. Journal of biomedical informatics 2015;56:87-93. doi: 10.1016/j.jbi.2015.05.001 [published Online First: 2015/05/20]
13. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res;0(0):0962280214558972. doi: doi:10.1177/0962280214558972
14. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., et al. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Medical decision making : an international journal of the Society for Medical Decision Making 2001;21(1):45-56. doi: 10.1177/0272989x0102100106 [published Online First: 2001/02/24]
15. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245-7. doi: 10.1016/j.jclinepi.2015.04.005 [published Online First: 2015/05/20]
16. Douglas VC, Hessler CS, Dhaliwal G, et al. The AWOL tool: derivation and validation of a delirium prediction rule. J Hosp Med 2013;8(9):493-9. doi: 10.1002/jhm.2062 [published Online First: 2013/08/08]
17. Freter SH, George J, Dunbar MJ, et al. Prediction of delirium in fractured neck of femur as part of routine preoperative nursing care. Age and ageing 2005;34(4):387-8. doi: 10.1093/ageing/afi099 [published Online First: 2005/06/16]
18. Inouye SK, Viscoli CM, Horwitz RI, et al. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. Annals of internal medicine 1993;119(6):474-81. [published Online First: 1993/09/15]
19. Inouye SK, Charpentier PA. Precipitating factors for delirium in hospitalized elderly persons. Predictive model and interrelationship with baseline vulnerability. Jama 1996;275(11):852-7. [published Online First: 1996/03/20]

20. Inouye SK, Zhang Y, Jones RN, et al. Risk factors for delirium at discharge: development and validation of a predictive model. Archives of internal medicine 2007;167(13):1406-13. doi: 10.1001/archinte.167.13.1406 [published Online First: 2007/07/11]
21. Carrasco MP, Villarroel L, Andrade M, et al. Development and validation of a delirium predictive score in older people. Age and ageing 2014;43(3):346-51. doi: 10.1093/ageing/aft141 [published Online First: 2013/09/26]
22. Rudolph JL, Harrington MB, Lucatorto MA, et al. Validation of a medical record-based delirium risk assessment. Journal of the American Geriatrics Society 2011;59 Suppl 2:S289-94. doi: 10.1111/j.1532-5415.2011.03677.x
23. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49(12):1373-9. [published Online First: 1996/12/01]
24. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS medicine 2014;11(10):e1001744. doi: 10.1371/journal.pmed.1001744 [published Online First: 2014/10/15]
25. de Wit HAJM, Winkens B, Mestres Gonzalvo C, et al. The development of an automated ward independent delirium risk prediction model. Int J Clin Pharm 2016 doi: 10.1007/s11096-016-0312-7
26. Leung JM, Sands LP, Lim E, et al. Does preoperative risk for delirium moderate the effects of postoperative pain and opiate use on postoperative delirium? The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry 2013;21(10):946-56. doi: 10.1016/j.jagp.2013.01.069 [published Online First: 2013/05/11]
27. Liang C-K, Chu C-L, Chou M-Y, et al. Developing a Prediction Model for Post-Operative Delirium and Long-Term Outcomes Among Older Patients Receiving Elective Orthopedic Surgery: A Prospective Cohort Study in Taiwan. Rejuvenation Res 2015;18(4):347-55. doi: 10.1089/rej.2014.1645
28. Rudolph JL, Jones RN, Levkoff SE, et al. Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. Circulation 2009;119(2):229-36 8p. doi: 10.1161/CIRCULATIONAHA.108.795260
29. Grobman WA, Stamilio DM. Methods of clinical prediction. American journal of obstetrics and gynecology 2006;194(3):888-94. doi: 10.1016/j.ajog.2005.09.002 [published Online First: 2006/03/09]
30. Subramanian J, Simon R. Overfitting in prediction models - is it a problem only in high dimensions? Contemporary clinical trials 2013;36(2):636-41. doi: 10.1016/j.cct.2013.06.011 [published Online First: 2013/07/03]
31. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol 2016;76:175-82. doi: 10.1016/j.jclinepi.2016.02.031 [published Online First: 2016/03/12]
32. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. BMJ (Clinical research ed) 2015;351:h3868. doi: 10.1136/bmj.h3868 [published Online First: 2015/08/13]
33. Dworkin A, Lee DS, An AR, et al. A Simple Tool to Predict Development of Delirium After Elective Surgery. Journal of the American Geriatrics Society 2016;64(11):e149-e53. doi: 10.1111/jgs.14428 [published Online First: 2016/09/22]
34. Fisher BW, Flowerdew G. A simple model for predicting postoperative delirium in older patients undergoing elective orthopedic surgery. Journal of the American Geriatrics Society 1995;43(2):175-8. [published Online First: 1995/02/01]
35. Korc-Grodzicki B, Sun SW, Zhou Q, et al. Geriatric Assessment as a Predictor of Delirium and Other Outcomes in Elderly Patients With Cancer. Annals of surgery 2014 doi: 10.1097/sla.0000000000000742 [published Online First: 2014/06/03]
36. O'Keeffe ST, Lavan JN. Predicting delirium in elderly patients: development and validation of a risk-stratification model. Age and ageing 1996;25(4):317-21. [published Online First: 1996/07/01]
37. Pompei P, Foreman M, Rudberg MA, et al. Delirium in hospitalized older persons: outcomes and predictors. Journal of the American Geriatrics Society 1994;42(8):809-15. [published Online First: 1994/08/01]

38. Kim MY, Park UJ, Kim HT, et al. DELirium Prediction Based on Hospital Information (Delphi) in General Surgery Patients. Medicine (Baltimore) 2016;95(12):e3072. doi: 10.1097/MD.0000000000003072
39. Kalisvaart KJ, Vreeswijk R, de Jonghe JF, et al. Risk factors and prediction of postoperative delirium in elderly hip-surgery patients: implementation and validation of a medical risk factor model. Journal of the American Geriatrics Society 2006;54(5):817-22. doi: 10.1111/j.1532-5415.2006.00704.x [published Online First: 2006/05/16]
40. Pendlebury ST, Lovett NG, Smith SC, et al. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of a susceptibility score based on factors identified externally in pooled data for use at entry to the acute care pathway. Age and ageing 2016 doi: 10.1093/ageing/afw198 [published Online First: 2016/11/07]
41. Moerman S, Tuinebreijer WE, de Boo M, et al. Validation of the Risk Model for Delirium in hip fracture patients. Gen Hosp Psychiatry 2012;34(2):153-9. doi: 10.1016/j.genhosppsych.2011.11.011 [published Online First: 2012/01/10]
42. Freter S, Dunbar M, Koller K, et al. Risk of Pre-and Post-Operative Delirium and the Delirium Elderly At Risk (DEAR) Tool in Hip Fracture Patients. Can Geriatr J 2015;18(4):212-16. doi: 10.5770/cgj.18.185
43. Pendlebury ST, Lovett N, Smith SC, et al. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of externally derived risk scores. Age and ageing 2016;45(1):60-65. doi: 10.1093/ageing/afv177
44. Rudolph JL, Doherty K, Kelly B, et al. Validation of a Delirium Risk Assessment Using Electronic Medical Record Information. Journal of the American Medical Directors Association 2016;17(3):244-48 5p. doi: 10.1016/j.jamda.2015.10.020
45. Freter SH, Dunbar MJ, MacLeod H, et al. Predicting post-operative delirium in elective orthopaedic patients: the Delirium Elderly At-Risk (DEAR) instrument. Age and ageing 2005;34(2):169-71.
46. Martinez JA, Belastegui A, Basabe I, et al. Derivation and validation of a clinical prediction rule for delirium in patients admitted to a medical ward: an observational study. BMJ Open 2012;2(5) doi: 10.1136/bmjopen-2012-001599 [published Online First: 2012/09/18]
47. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010;172(8):971-80. doi: 10.1093/aje/kwq223 [published Online First: 2010/09/03]
48. Isfandiaty R, Harimurti K, Setiati S, et al. Incidence and predictors for delirium in hospitalized elderly patients: a retrospective cohort study. Acta medica Indonesiana 2012;44(4):290-7. [published Online First: 2013/01/15]

VERSION 2 – REVIEW

REVIEWER	Sarah T Pendlebury Centre for Prevention of Stroke and Dementia, Nuffield Department of Clinical Neurosciences, University of Oxford and NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford
REVIEW RETURNED	05-Feb-2018

GENERAL COMMENTS	The authors have responded to the reviewers' comments in detail. I have only a few minor comments: Results, second paragraph, reference 50 is missing from the list of externally validated models. Table 1. Why are the Newcastle Ottawa quality scores different for the two Pendlebury studies when these used the same cohort? Also, why is the comparability score <2 when this was a consecutive

	cohort of acute medicine patients without any exclusion criteria? For the Pendlebury study ref 50, this should be listed in the table as "Val" rather than "Dev" as it is a validation study . Results. Ref 50 should be added to the list of models whose variable selection was made using variables based on a literature review. Results, Variables, Functional Impairment. In studies 34 and 50, functional impairment was defined as needing a care package (professional carers at home) or residence in a care home since these details were easily ascertained on admission – it would be good to add this to the text here.

REVIEWER	Richard N. Jones Brown University, USA
REVIEW RETURNED	07-Feb-2018

GENERAL COMMENTS	Thank you for responding to my prior comments. I have no new comments.

REVIEWER	Diether Kramer Steiermärkische Krankenanstaltengesellschaft m.b.H., Austria
REVIEW RETURNED	11-Feb-2018

GENERAL COMMENTS	The authors have taken into account the suggestions respectively provided comprehensible reasons not to do so completely.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 3
Reviewer Name: Sarah T Pendlebury
Institution and Country: Centre for Prevention of Stroke and Dementia, Nuffield Department of Clinical Neurosciences, University of Oxford and NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford
Please state any competing interests: I am lead author on two of the cited studies (34,50)

Please leave your comments for the authors below

1. The authors have responded to the reviewers' comments in detail. I have only a few minor comments:

Response: Thank you for your detailed feedback. We are glad you are pleased with our responses.

2. Results, second paragraph, reference 50 is missing from the list of externally validated models.

Response: This is corrected.

3. Table 1. Why are the Newcastle Ottawa quality scores different for the two Pendlebury studies when these used the same cohort? Also, why is the comparability score <2 when this was a consecutive cohort of acute medicine patients without any exclusion criteria?

Response: This is a valued question and the authors have re-evaluated both studies. The Newcastle Ottawa quality scores are different for the two Pendlebury studies because the criteria for comparability are the following:

“Comparability of cohorts on the basis of the design or analysis:
a) Study controls for ___ (select the most important factor) *
b) Study controls for any additional factor (this cirteria could be modified to indicate specific control for a second important factor)*”

In one study, the models were stratified by age, but odds ratios or models were not adjusted for covariates and baseline factors. In the second study, the models were adjusted for age and models were weighted for each included factor. We have adjusted the NOS score for the second study from 8 to 9 for this reason. On page 7, the number of studies receiving a maximum of nine stars was changed from five to six

“The average NOS quality ranking for included cohort studies was seven; six studies received the maximum of nine stars.”

4. For the Pendlebury study ref 50, this should be listed in the table as “Val” rather than “Dev” as it is a validation study .

Response: This is corrected.

5. Results. Ref 50 should be added to the list of models whose variable selection was made using variables based on a literature review.

Response: This is corrected on pg 12.

6. Results, Variables, Functional Impairment. In studies 34 and 50, functional impairment was defined as needing a care package (professional carers at home) or residence in a care home since these details were easily ascertained on admission – it would be good to add this to the text here.

Response: The authors appreciate this feedback and it has been added to the paragraph on page 12. The updated paragraph is below:

“Functional impairment was defined as follows: (1) needing assistance with any basic ADL,27 (1) domestic help, help with meals or physical care41 and (2) residence in nursing facility or at home with caregivers33, (2) requiring a home care package with professional caregivers or residence in a care home.33 49 The latter being obtained upon admission from medical records.33 49 Two studies used validated functional assessment tools (iADL and Barthel Index) and evaluated functional status two weeks prior to hospitalization.23 31”

Reviewer: 4
Reviewer Name: Richard N. Jones
Institution and Country: Brown University, USA
Please state any competing interests: None declared

Please leave your comments for the authors below
1. Thank you for responding to my prior comments. I have no new comments.

Response: Thank you for your time and efforts.

Reviewer: 2
Reviewer Name: Diether Kramer
Institution and Country: Steiermärkische Krankenanstaltengesellschaft m.b.H., Austria
Please state any competing interests: None

Please leave your comments for the authors below
1. The authors have taken into account the suggestions respectively provided comprehensible reasons not to do so completely.

Response: Thank you for your feedback and work on this review.

VERSION 3 – REVIEW

REVIEWER	Sarah T Pendlebury Centre for Prevention of Stroke and Dementia, and the NIHR Oxford Biomedical Research Centre Univeristy of Oxford and the John Radcliffe Hospital, Oxford, UK
REVIEW RETURNED	15-Mar-2018

GENERAL COMMENTS	The authors have responded satisfactorily to the comments. I would like to commend the authors for their detailed and comprehensive responses to the reviewers throughout the process