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)	Predictive models for short- and long-term improvement in women under physiotherapy for chronic disabling neck pain: a longitudinal cohort study
AUTHORS	Bohman, Tony; Bottai, Matteo; Björklund, Martin

VERSION 1 – REVIEW

REVIEWER	Ney Meziat-Filho Centro Universitário Augusto Motta (UNISUAM), Brazil
REVIEW RETURNED	08-Jul-2018

GENERAL COMMENTS	Very relevant research question and a well-written paper. There are only two issues in my concern. 1 – Even considering the justifications of the authors, the sample size seems to be small for a prediction study and may have increased the risk of type 1 error. 2 –If the interventions used in the clinical trial from where the data were extracted approach the psychological (cognitive/affective) dimension of pain, would the results change? I suggest a discussion about the next steps for the validation of the predictive model, including how the clinical utility would be tested.

REVIEWER	Ingebrigt Meisingset Norwegian University of Science and Technology, Norway
REVIEW RETURNED	04-Sep-2018

GENERAL COMMENTS	The study is a secondary analyses of a RCT developing prognostic models for clinically important improvement in subjects with neck pain. The study thus concerns an important topic in the musculoskeletal field. I have several concerns regarding the manuscript and these needs to be addressed before the review process continue. Minor revisions are therefore not commented in this stage of the process. My main concerns are: 1. Using GEE as statistical analysis provide population-specific estimates, as correctly specified by the authors. However, prediction models are commonly used to infer prognosis on an individual level and the analysis should thus use subject specific analysis, such as mixed models. The discussion (page 22, line 33-39) discusses the results and infers on an individual level (clinically). An other important advantage with using mixed models is that GEE has an assumption of missing data completely at

	random, while mixed models allows for missing data under a weaker assumption, missing at random. The first is probably not true for this study. What was the reason for choosing a population specific analysis in developing a prognostic model? 2. The selection of potential variables are not clearly described, as the reader is unable to decide what variables were chosen based on previous research or by clinical considerations and availability in the data (page 7,line 34-38). Further, it is unclear if the references for each potential variable in Table 1 is related to the prognostic ability of the variable or to information on the questionnaire used. This information needs to be specified in order to understand how the variables were chosen. 3. The data used for the development of the prediction models come from a RCT. This is suitable, but it is important that the variables are adjusted for the treatment. This is done in other studies using data from RCT (for example Schellingerhout et al. 2010). The authors should the treatment variable as an covariate. 3. Some variables from the original RCT study were not included as potential prognostic variables, even though they have been shown to predict outcome in other studies (e.g. duration of current pain episode) . Why was this variable not included? 4. The language could benefit from proof reading by a native English speaking person . There are too many long sentences and the language could have been more direct (e.g. page 19,line 30-34, page 20,line 20-24, and page 21 line 23-29). 5.Page 11,line 48: Here it is stated that the complete data sets. This is somewhat confusing as this can be interpreted as the analysis only included subjects with data at all time points. 6. Descriptives are presented as median (ICR). Is it due to non-normal distribution on the continuous variables? Can the mean (SD) be provided where there is normally distributed data? 7. According to the PROGRESS guidelines and other publications , internal validation of prediction models should include bootstrapping techniques, as provided in the current manuscript, followed by an adjustment for optimisms (shrinkage methods).The latter is not performed in this study. The authors should consider using shrinkage as the estimates from the analysis are likely to be too optimistic (see ref PROGRESS for developing prognostic models and Steyerberg et al. 2004 (https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1844). 7. I would like to have the 95 % CI in addition to the p-value in table 3 for the univariable analyses instead to be able to judge the precision of the estimates. 8. Page 22, line 25. It is stated that the results are in with the study of Bot et al. However, Bot et al. found that baseline disability could predict functional disability but not recovery. This needs to be rewritten. In the previuos sentence, line 23, it is stated that baseline disability level were associated with recovery in both models. However,one model used recovery as outcome while the other model used clinically important improvement as outcome. These outcomes are different as a patient may have a clinically important

	improvement but still not report recovered. The authors must distinguish between these outcomes. 9. I miss some references in the manuscript, for example page 22, line 46-48 Minor: 1. Page 22,line 2-15 discuss the influence of attrition and that patients seeking health care might be different from those who were included by advertisement. A study by our group (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782331/pdf/12913_2016_Article_1326.pdf) suggest that people seeking treatment are different from those who do not seek treatment, i.e. they have poorer self-reported health, symptoms of anxiety or depression, they are more obese and smoke more often, and have coexisting neck and low back pain. The results from the abovementioned study should be discussed.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Ney Meziat-Filho
Institution and Country: Centro Universitário Augusto Motta (UNISUAM), Brazil Please state any competing interests or state 'None declared': None declared

Very relevant research question and a well-written paper. There are only two issues in my concern.

1 – Even considering the justifications of the authors, the sample size seems to be small for a prediction study and may have increased the risk of type 1 error.
Response:
As, mentioned in the discussion section of the manuscript, we do have some concerns about the small sample size. It may increase the risk of type 1 error as well as the risk of over dispersion. However, we have reasons to believe that the associations we observed can be explained by priori knowledge of prognostic factors for neck pain, and are supported by similar results found in other studies concerning such factors. We believe this would augment the results of our statistical tests and lend them support even with this relatively small sample.
Furthermore, we added the calculated shrinkage factor as a valuation of over dispersion in the revised manuscript (statistical method, page 13, lines 1-2 and results, page 16, line 10-11).

2 –If the interventions used in the clinical trial from where the data were extracted approach the psychological (cognitive/affective) dimension of pain, would the results change? I suggest a discussion about the next steps for the validation of the predictive model, including how the clinical utility would be tested.
Response:
The result of our study is confined to our study population and the type of treatment interventions investigated, both which are resembling the situation in primary care. However, to clarify the next step of validation of the predictive model and its clinical utility, we have added a more detailed description of this at the end of the discussion section (see page 21, line 13-16).

Reviewer: 2

Reviewer Name: Ingebrigt Meisingset
Institution and Country: Norwegian University of Science and Technology, Norway Please state any competing interests or state 'None declared': None declared

The study is a secondary analyses of a RCT developing prognostic models for clinically important improvement in subjects with neck pain. The study thus concerns an important topic in the musculoskeletal field.

I have several concerns regarding the manuscript and these needs to be addressed before the review process continue. Minor revisions are therefore not commented in this stage of the process.

My main concerns are:
1. Using GEE as statistical analysis provide population-specific estimates, as correctly specified by the authors. However, prediction models are commonly used to infer prognosis on an individual level and the analysis should thus use subject specific analysis, such as mixed models. The discussion (page 22, line 33-39) discusses the results and infers on an individual level (clinically). An other important advantage with using mixed models is that GEE has an assumption of missing data completely at random, while mixed models allows for missing data under a weaker assumption, missing at random. The first is probably not true for this study. What was the reason for choosing a population specific analysis in developing a prognostic model?
Response:
Thank you for a very relevant comment. As a prediction model is meant to help clinicians in their management of individuals mixed models would be a better choice than GEE. Therefore, we have reanalysed the data using Random-effects logistic regression models. This resulted in a final model including the same predictors as when using GEE analyses when NDI was the outcome. However, the change of analyses method resulted in inclusion of age instead of recovery expectations in the final model with PGICS as outcome. Text, as well as tables, have been rewritten and corrected in "Statistical methods", "Results" and "Discussion" according to the change of statistical method. We hope you find these changes satisfactory.

2. The selection of potential variables are not clearly described, as the reader is unable to decide what variables were chosen based on previous research or by clinical considerations and availability in the data (page 7, line 34-38). Further, it is unclear if the references for each potential variable in Table 1 is related to the prognostic ability of the variable or to information on the questionnaire used. This information needs to be specified in order to understand how the variables were chosen.
Response:
Thank you for making us aware of the confusion. To clarify we have included additional text explaining which of the potential prognostic factors were chosen based on prior knowledge and which ones were chosen according to clinical considerations, page 7, line 17-22. We have also explained that reference included in table 1 refers to the definition and to psychometric properties of the potential prognostic factors in the table 1 header.

3. The data used for the development of the prediction models come from a RCT. This is suitable, but it is important that the variables are adjusted for the treatment. This is done in other studies using data from RCT (for example Schellingerhout et al. 2010). The authors should the treatment variable as an covariate.
Response:
The reason for not adjusting for the treatment in our analyses was that, when the study is based on an RCT, we believe that the potential prognostic factors (predictors) were randomly distributed between treatments and therefore adjusting for treatment were unnecessary. But when, as you mention, adjusting for treatment is common in other studies we have followed your recommendation

and adjusted for the treatment in the multivariable analyses. This is explained tin the “Statistical methods”, page 12, line 17-19.

3. Some variables from the original RCT study were not included as potential prognostic variables, even though they have been shown to predict outcome in other studies (e.g. duration of current pain episode) . Why was this variable not included?
Response:
It is true that longer duration of pain could be a potential prognostic variable, however, this was mainly shown for comparisons between acute versus subacute, versus chronic pain [1-4]. In our sample, all participants reported neck pain duration of more than eight months, with a median duration of 120 months (IQR 60-216). Thus, our participants had very long-term pain and all belonged, with margin, to the group chronic pain. Michaelson et. al found no association between baseline duration of pain (mean duration; 106 months) and short- or long-term pain intensity following chronic neck pain patients in multimodal rehabilitation [5]. There is, to our knowledge, no scientific support for that different pain duration within the category chronic pain has any prognostic relevance, which is why we did not include pain duration in the analysis.
1. Bot, S.D.M., et al., Predictors of outcome in neck and shoulder symptoms - A cohort study in general practice. Spine, 2005. 30(16): p. E459-E470.
2. Feleus, A., et al., Prognostic indicators for non-recovery of non-traumatic complaints at arm, neck and shoulder in general practice-6 months follow-up. Rheumatology, 2007. 46(1): p. 169-176.
3. Karels, C.H., et al., Social and psychological factors influenced the course of arm, neck and shoulder complaints. Journal of Clinical Epidemiology, 2007. 60(8): p. 839-848.
4. Keijsers, E., et al., Psychosocial factors predicted nonrecovery in both specific and nonspecific diagnoses at arm, neck, and shoulder. Journal of Clinical Epidemiology, 2010. 63(12): p. 1370-1379.
5. Michaelson P, Sjolander P, Johansson H. Factors Predicting Pain Reduction in Chronic Back and Neck Pain After Multimodal Treatment. The Clinical Journal of Pain 2004;20(6).

4. The language could benefit from proof reading by a native English speaking person . There are too many long sentences and the language could have been more direct (e.g. page 19,line 30-34, page 20,line 20-24, and page 21 line 23-29).
Response:
The co-author Matte Bottai is a native English speaking person and he has now thoroughly checked the language. We sincerely hope you will find the language to be acceptable in the revised manuscript.

5.Page 11,line 48: Here it is stated that the complete data sets. This is somewhat confusing as this can be interpreted as the analysis only included subjects with data at all time points.
Response:
Yes, you are right, it is confusing. The analyses included subjects with follow-up data for at least one follow-up. We have simply deleted the sentence as it does not add any relevant information. We have already explained how the study population was determined in the first paragraph under “Data collection and variables”, page 6, line 18-19, why the sentence is redundant.

6.
Descriptives are presented as median (ICR). Is it due to non-normal distribution on the continuous variables? Can the mean (SD) be provided where there is normally distributed data?
Response:
As only two of the continuous variables were normally distributed we decided
to just report the median and IQR. According to your suggestions, we have changed the text in “Statistical methods”, page 11, line 11-12, and included means and SD for the normally distributed variables (average pain intensity and social support) in table 2.

7. According to the PROGRESS guidelines and other publications , internal validation of prediction models should include bootstrapping techniques, as provided in the current manuscript, followed by an adjustment for optimisms (shrinkage methods).The latter is not performed in this study. The authors should consider using shrinkage as the estimates from the analysis are likely to be too optimistic (see ref PROGRESS for developing prognostic models and Steyerberg et al. 2004 (https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1844).
Response:
We agree, including an adjustment for optimism is valuable. Thus, in the revised manuscript the heuristic shrinkage factor calculated according to Steyerberg et al. was used as adjustment for optimism. Information referring to the shrinkage factor calculation is included in “Statistical methods” (page 13, line 1-2), and in the results, page 16, line 10-11. Steyerberger et al. 2004 is included as reference (#30) in “Statistical methods”.

7. I would like to have the 95 % CI in addition to the p-value in table 3 for the univariable analyses instead to be able to judge the precision of the estimates.
Response:
We have included the 95% CI in table 3 as requested.

8. Page 22, line 25. It is stated that the results are in with the study of Bot et al. However, Bot et al. found that baseline disability could predict functional disability but not recovery. This needs to be rewritten. In the previuos sentence, line 23, it is stated that baseline disability level were associated with recovery in both models. However,one model used recovery as outcome while the other model used clinically important improvement as outcome. These outcomes are different as a patient may have a clinically important improvement but still not report recovered. The authors must distinguish between these outcomes.
Response:
Thank you for the comment. We have corrected our text related to the findings of Bot et al. in accordance with your remark, page 20, line 23-25.
Both our outcomes (PGICS and NDI) are based on clinical important improvement. We apologies for not properly distinguish between clinical important improvement, clinical important change and recovery in the original manuscript. This has led to confusion. In the revised manuscript we have corrected these terms in relation to our outcome, clinical important improvement. Therefore, several changes are made in the method and discussion parts. We have not changed these terms when they have been used in other original studies, as for example in the introduction.

9. I miss some references in the manuscript, for example page 22, line 46-48 – tobacco use etc –
Response:
We added two references and made some small changes in the text to further clarify, page 21, line 6-8.

Minor:
1. Page 22,line 2-15 discuss the influence of attrition and that patients seeking health care might be different from those who were included by advertisement. A study by our group (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782331/pdf/12913_2016_Article_1326.pdf) suggest that people seeking treatment are different from those who do not seek treatment, i.e. they have poorer self-reported health, symptoms of anxiety or depression, they are more obese and smoke more often, and have coexisting neck and low back pain. The results from the abovementioned study should be discussed.
Response:

Thank you for your reference. It is a valuable contribution to our statement. We have considered your suggestion to discuss the article. However, as we already have elaborated on this issue, we decided just to include the reference, page 20, line 16-17.

VERSION 2 – REVIEW

REVIEWER	Ingebrigt Meisingset Norwegian University of Science and Technology
REVIEW RETURNED	14-Nov-2018

GENERAL COMMENTS	Thanks for considering my comments and the revisions done. The paper is improved. However, there are still several major concerns related to the methods used in the paper. I believe this must be addressed in order to improve the quality of the article and the general scientific evidence and quality concerning prognostic models in patients with musculoskeletal pain. My comments are related to the manuscript with track changes (Page 33-79) Major • It is not clear why only results for 3 and 15 months are reported while the analysis also included outcome at 9 months. The 95% CI are thus dependent on all the 3 follow ups. This must be clarified in the manuscript and be reported consistently according to the analyses performed. • In predictive models and calculation of sample size it is the number of subjects with the outcome event(clinically important improvement) that is important, not the number of outcome observations. See TRIPOD (point 14A) for elaboration. This must be rewritten in the manuscript and emphasized in the limitation section. The tables must decribe number of outcome events in addition to number of outcome observations. • There is some uncertainty whether the results presented are estimates for the outcomes at 3,9 and 15 months follow up or is limited to 3 and 15 months follow up. This is due to inconsistency in the description of the analyses in the methods section, and in the description of the results. E.g the interaction terms do not include 9 months while the number of outcome observations include 9 months. This must be clarified for the reader throughout the manuscript. • Page 50 line 42-: How was the shrinkage factor used to address optimism of the estimates in the prediction model? The shrinkage factor must be used to reduce overfitting (shrinkage factor Xestimates) especially in small sample size studies such as the current study, not only reporting the value. See TRIPOD and PROGRESS articles for elaboration. • Page 63 line3-: The authors shopuld elaborate more the results for catastrophizing and provide some suggestions for why this results were observed and how clinicians should interpret the findings, as the authors state that clinicians should notice this in clinical management.

	• Page 65 line 8-19: This section must be rewritten according the comments above. Sample size is not calculated based on number of variables in the analysis, but number of participants per parameter estimated (e.g. an interaction term provide 2 parameter estimates and a categorical variable provides one parameter per category). Including interaction terms with time reduces power substantially. The number of parameters are thus larger than the authors describe in the article. This needs to be addressed in the method section, the results and discussed in the limitation section. Minor: • Page 36 line 17-: The abstract should describe the direction of the association between the predictors and the outcome, e.g. lower age and higher neck disability predicted improvement… • Page 36 line 50: Delete possible as the small sample size is clearly a limitation in this study. • Page 37 first paragraph (introduction): There is a lot of information on prevalence and burden. This section could be easier to read if it is shortened (too many numbers). Consider revising this section. • Page 41 line 18-: How many subjects had a NDI% of 0 at follow up? This is important to know for the reader and is not reported in the manuscript. • Page 50 line 14-: What was defined as signs of collinearity in the analysis? This must be described. • Univariate and multivariable analysis are used as terms. The correct terms when analyzing longitudinal data with several time points is univariate and multivariate. Please be consistent throughout the manuscript. • According to TRIPOD, the intercept in a prediction final model must be reported. Otherwise, it is imposible for the reader to use the prediction model. Please report the intercept. • In the authors response to my previuos comments, the authors argue for the exclusion of pain duration as a predictor. However, I suggest an sensitivity analysis including pain duration should have been performed instead of arguing to not include pain duration in the univariate analysis. • Page 67 line21: Delete “the chance of”

VERSION 2 – AUTHOR RESPONSE

Dear reviewer,

Thank you for your valuable comments! We have considered your comments and made changes accordingly.

Please, find our response to your comments below. Our responses are related to the “Main document – marked copy”.

Major

• It is not clear why only results for 3 and 15 months are reported while the analysis also included outcome at 9 months. The 95% CI are thus dependent on all the 3 follow ups. This must be clarified in the manuscript and be reported consistently according to the analyses performed.

Response:

Our aim of the study was to develop predictive models for short and long term effects, ie at remission from intervention (3-month follow-up) and at one year after remission (15-month follow-up). The 9-month follow-up (ie 6 months after remission) we consider an intermediate follow-up (1) and we had no intention to report this even from the beginning. Therefore, we decide to stick to our pre-determined aim and not include the 9-month follow-up in eg the result tables. As a consequence of our decision, we have only reported the estimates for catastrophizing at 3 and 15 months in table 3 and 4.

However, in order to further clarify that the 9-month outcome also is included in the analyses we have added some information at page 11, line 17, and at page 12, line one. In addition, we have included a short clarification in the footnotes in table 3 and 4. We hope you find our clarifications satisfactory.

As to your comment about the 95% CI we are not sure what you mean, but as you rightly state, the estimates in a longitudinal analysis are dependent on all outcomes taken on each participant. We hope that our above mentioned changes together with our already included statement at page 11, line 17 to 19; “Random-effects logistic regression can model longitudinal data efficiently, while taking into account the potential dependence of the repeated measures taken on each participant.” will make this clear.

1. SBU. Rehabilitering vid långvarig smärta. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 198. ISBN 978-91-85413-34-8. Kapitel 3.2, Inklusionskriterier. In Swedish only.

• In predictive models and calculation of sample size it is the number of subjects with the outcome event(clinically important improvement) that is important, not the number of outcome observations. See TRIPOD (point 14A) for elaboration. This must be rewritten in the manuscript and emphasized in the limitation section. The tables must decribe number of outcome events in addition to number of outcome observations.

Response:

Thank you for raising this point as we may have been unclear. This study does not have any censored observations why we do not have “events”. Our outcome is, as you rightly address, clinical important improvement, and we have now included information about the number of clinical improved cases at each follow-up in the result section, page 13, line 13 to 16. We believe, reporting cases at all three follow-ups for each analysis in the tables will just be confusing for the reader and a matter of over reporting. We hope you find our decision acceptable.

Including number of observations in the tables gives important information about internal missing as some participants did not report on all three follow-ups, and we still think this should be reported along with the number of participants in the tables.

However, we have omitted the paragraph discussing numbers of observations, page 21, line 9-10, as the number of observations is not directly related to the power.

• There is some uncertainty whether the results presented are estimates for the outcomes at 3,9 and 15 months follow up or is limited to 3 and 15 months follow up. This is due to inconsistency in the description of the analyses in the methods section, and in the description of the results. E.g the interaction terms do not include 9 months while the number of outcome observations include 9 months. This must be clarified for the reader throughout the manuscript.

Response:

It is unfortunate that you find our methods and result section to be unclear in this matter, our apologies. We hope that our changes explained above (page 11, line 17, page 12, line 1, and table 3 and 4) will further clarify. The analyses including the interaction were also based on data from all three follow-ups, but only reported for 3 and 15 months as it was our predetermined aim of the study.

We have also changed the text at page 12 line 9 to emphasise that the interaction analyses with time and the "check" for effect measure modification were two different analyses as this may have led to confusion in the former manuscript.

- Page 50 line 42-: How was the shrinkage factor used to address optimism of the estimates in the prediction model? The shrinkage factor must be used to reduce overfitting (shrinkage factor Xestimates) especially in small sample size studies such as the current study, not only reporting the value. See TRIPOD and PROGRESS articles for elaboration.

Response:
We were not able find any information in TRIPOD or PROGRESS saying that the estimates should be presented corrected by the shrinkage. However, due to your recommendation we added some text in the method section, page 13, line 3-4, and a column in table 4 with the shrinkage corrected beta-coefficients (Sβ).

- Page 63 line3-: The authors shopuld elaborate more the results for catastrophizing and provide some suggestions for why this results were observed and how clinicians should interpret the findings, as the authors state that clinicians should notice this in clinical management.

Response:
Thank you for your comment. As you say, we have already discussed that the odds of improvement changes over time for catastrophizing at page 19, lines 15-18, and prefer to keep it that way. However, due to your comment, we have added a sentence at page 22, line 10-12 to further elaborate about the possible use of our results for clinicians.
Important though, is to stress that as this study only includes the development of predictive models the clinical interpretation of the results should be considered with great caution, something we discussed at page 19, line 21-22.
In the end of the discussion we also mentioned that to be of clinical use the models has to be externally validated. This is in accordance with the PROGRESS recommendations; "Claiming that a model is clinically valuable is acceptable only with an external validation study using independent data from a different location than the development data" (1). Therefore, we have been deliberately conservative in giving recommendation on how clinicians should use the results.

A discussion of why the result for catastrophizing were observed would only be hypothetical and not fully supported by our study as the models are not externally validated. Hypothetically, we believe it reasonable that persons with high levels of catastrophizing have a good chance of improving at short term after physiotherapy, and that this effect will diminish by time as their tendency to catastrophize will affect their self-perceived improvement negatively. However, we hope you find it okay not to include this hypothesis in the manuscript due to our explanation above.

1. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.

- Page 65 line 8-19: This section must be rewritten according the comments above. Sample size is not calculated based on number of variables in the analysis, but number of participants per parameter estimated (e.g. an interaction term provide 2 parameter estimates and a categorical variable provides one parameter per category). Including interaction terms with time reduces power substantially. The number of parameters are thus larger than the authors describe in the article. This needs to be addressed in the method section, the results and discussed in the limitation section.

Response:
You are perfectly right, the sample size should be discussed in accordance with the number of parameters (coefficients) and not the number of variables. Accordantly, we have changed the text in the discussion, page 21, line 8. We have also emphasised the small sample as a limitation, page 21, line 5.
In the first step of the backward predictive model analyses with NDI as outcome we had 88 participants and 9 coefficients; one each for NDI, age, depression, and physical activity, two for intervention and three for catastrophizing with the time-interaction. In all other analyses the number of parameters were smaller. Thus we believe our statement in the manuscript, page

21, line 7-9, about our sample size to be large enough to ensure the recommendations ("rule of thumb") of 10 to 15 participants for each parameter estimate (coefficient) still to be valid. Furthermore, we have deleted the following sentence (page 21, line 9-10) about the number of observations supporting our sample size as it was confusing. We have also followed your recommendations below and deleted "possible" at page 3, line 20.

Minor:

- Page 36 line 17-: The abstract should describe the direction of the association between the predictors and the outcome, e.g. lower age and higher neck disability predicted improvement…

  Response:
  We believe the main results of or study to be the identification of the potential prognostic factors that best predicted the outcomes, and that the predictive models had an acceptable predictive ability. As the models has not yet been externally validated we have been deliberately cautious when discussing the use in clinic and also the direction of the associations as well as the magnitude of the estimates. But as we shortly elaborated on the direction of the associations in the discussion, we have followed your advice and included a description of the direction of associations in the abstract as well.

- Page 36 line 50: Delete possible as the small sample size is clearly a limitation in this study.

  Response:
  Done, page 3, line 20.

- Page 37 first paragraph (introduction): There is a lot of information on prevalence and burden. This section could be easier to read if it is shortened (too many numbers). Consider revising this section.

  Response:
  We have followed your recommendation and revised the section, page 4, line 3-10.

- Page 41 line 18-: How many subjects had a NDI% of 0 at follow up? This is important to know for the reader and is not reported in the manuscript.

  Response:
  We have included the information at page 13, line 16-18.

- Page 50 line 14-: What was defined as signs of collinearity in the analysis? This must be described.

  Response:
  We followed standard procedures and looked for signs of collinearity during the sequential backward manual selection, eg estimates or standard error with extremely high values or remarkable changes during the deletion process. We found no such signs of collinearity. However, as we made no formal test to check collinearity our statements in the former manuscript may be misleading. Therefore, we decided to omit these statements from the method section, page 12, line 16-17, and from the result section, page 17, line 10.

- Univariate and multivariable analysis are used as terms. The correct terms when analyzing longitudinal data with several time points is univariate and multivariate. Please be consistent throughout the manuscript.

  Response:
  Thank you for raising this relevant and interesting question. We have thoroughly looked for a consensus for such terms in these types of analyses without success. For example, we were not able to find a consensus about how to use terms like multivariate (1) or multivariable (2, 3) in a longitudinal analysis. Thus, we decided to omit all such nomenclature (univariate and multivariable) throughout the manuscript in order not to confuse the reader.

1. Hidalgo B, Goodman M. Multivariate or Multivariable Regression? American Journal of Public Health. 2012;103(1):39-40.
2. Fitzmaurice GM. Applied longitudinal analysis. Laird NM, Ware JH, editors. Hoboken, N.J.: Hoboken, N.J. : Wiley-Interscience; 2004.
3. Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: a review. Statistical methods in medical research. 2014;23(1):42-59.

- According to TRIPOD, the intercept in a prediction final model must be reported. Otherwise, it is imposible for the reader to use the prediction model. Please report the intercept.

  Response:

  We report the intercepts in the footnote to table 4.

  However, we again want to emphasise what we have stated above and in the manuscript; we have just performed the first step (development) in presenting predictive models, meaning that it is too early to use the models in clinic as they have not been externally validated.

- In the authors response to my previuos comments, the authors argue for the exclusion of pain duration as a predictor. However, I suggest an sensitivity analysis including pain duration should have been performed instead of arguing to not include pain duration in the univariate analysis.

  Response:

  One basic concern in prediction modelling is to be careful when including factors to minimise the risk of findings by chance (ie increase the risk of type I error). Our decision not to include pain duration as a potential factor was based on the explanation given in our previous review answer and the fact that pain duration ranged from 8 to 456 months (median 120 months) in the study sample. We strongly believe that when reporting pain duration that long, the risk of information bias is obvious and possibly resulting in unreliable data.

  Following the explanation above we find it inappropriate and not common procedure to perform sensitivity analyses on factors pre-determined to be excluded, and therefore we refrain from undertaking a sensitivity analyses including pain duration. We hope you find our explanation reasonable.

- Page 67 line21: Delete “the chance of”

  Response:

  Done (page 22, line 25). Thank you for making us aware of this mistake!

VERSION 3 – REVIEW

REVIEWER	Ingebrigt Meisingset Norwegian University of Science and Technology
REVIEW RETURNED	24-Jan-2019

GENERAL COMMENTS	Thanks for the detailed response and clarification throughout the manuscript. I believe all my points are addressed either in the manuscript or in the response to my review. This have improved the manuscript substantially and the paper will contribute to the existing knowledge regarding prognosis and prediction models in neck pain.