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)	External validation of the Probability of Repeated admission (Pra) risk prediction tool in older community-dwelling people attending general practice: a prospective cohort study
AUTHORS	Wallace, Emma; McDowell, Ronald; Bennett, Kathleen; Fahey, Tom; Smith, Susan

VERSION 1 - REVIEW

REVIEWER	Mark Ebell MD, MS University of Georgia, United States
REVIEW RETURNED	13-May-2016

GENERAL COMMENTS	Background Nice overview. Methods It isn't clear until Table 1 that the information is being pulled from the GP Record. Or is it? In the text, page 5 line 14 it states "The Pra score was calculated using postal questionnaire data collected in 2010." Yet, Table 1 seems to state that almost all of the data comes from the GP record (which is preferable). Please clarify. Results Page 9, lines 33-39 and Figures 1 and 2: The ROC curves do not resemble the typical ROC curve, which is fitted more closely to the actual points. Using this kind of a "stairstep" approach is overly conservative and underestimates the area under the curve. Also, why only use the 3 pre-specified points? You could use cutpoints of 0.1, 0.2, 0.3, etc to calculate a more accurate curve. Table 4: Please report likelihood ratios and predictive values for each stratum, rather than a series of sensitivity/specificity pairs. The whole point of creating risk groups like this is to have richer information about each stratum, which is lost when you simply dichotomize. Obviously, there is a calibration problem, with many fewer patients admitted now that 10-20 years ago. That is not terribly surprising. So why don't you explore how different cutoffs perform, to inform future validation studies? For example, perhaps a moderate risk group of .35 to 0.50, high risk of 0.5 to 0.7, and very high risk of > 0.7? It is OK to do this kind of exploratory analysis if you are clear that the intent is to inform future validation studies. Finally, why not use deciles to create a true calibration curve? This is a post-estimation option in stata, using the HL command with the graph option.

	Discussion I think you should discuss overall changes in hospital admission rates. Are more patients being managed as outpatients who would have been admitted previously? That would explain the calibration problem. Perhaps you could speculate about an enhanced version of the rule that incorporates the local rates of hospital admission to better calibrate the Pra score for a particular community and healthcare system.

REVIEWER	Rónán O'Caoimh National University of Ireland, Galway
REVIEW RETURNED	03-Jun-2016

GENERAL COMMENTS	External validation of the Probability of Repeated admission (Pra) risk prediction tool in older community-dwelling people: a prospective cohort study Summary This paper presents an external predictive validation of the Probability of Repeated admission (Pra) risk-prediction tool against a modified version of the same instrument in an Irish primary care setting. The results show that both instruments have poor accuracy in predicting hospitalisation, consistent with similar research with the Pra and other short instruments that measure risk of hospitalisation. I am unconvinced about the utility of the Pra based on these results and disagree with the authors' optimistic view that this instrument could be used to target patients for inclusion in future trials to tackle ED admissions. That it has not been recently externally validated (>10 years as the authors highlight) might relate to the poor accuracy (that the authors themselves found in their own systematic review and was found by Kansagara et al 2011 and O'Caoimh et al 2015. Nevertheless, the study is well conducted and presented. Abstract In the conclusion the authors state that the Pra shows modest discrimination. While, this may represent a synthesis of the accuracy and perceived utility of the Pra, the c-statistics presented suggest that the accuracy is poor. The final conclusion should, in my opinion, be toned down to reflect this. The results/findings that are important should be stressed, i.e. 1) The modified version is not significantly more accurate than the original Pra but is less calibrated to predict hospitalisation (which is useful to establish). However, this data could be found from registry data (GP records in this case) making it easier to use than the original in clinical practice. 2) Both have poor accuracy similar to other short-risk prediction instruments used for this purpose. 3) Given the brevity of these instrument and comparability to more detailed and less accessible registry based algorithms, it may be a reasonable choice if a someone wished to use such an instrument. However, based on the psychometric properties presented here there is no convincing evidence that such an instrument will pay off the time etc. invested to use it, particularly in Irish General Practices, where time is limited. Strengths and limitations This same criticism is applied to the strengths and limitation

	summary. This is overly positive and does not fully reflect the results. The poor accuracy should be stated. See above points 1-3. Background Well presented. A minor point: in the 2nd paragraph a conjunction is missing “of the Pra tool AND identified 10....” Methods Please clarify the sample. Are these practices urban-suburban-rural or a mix? How representative of the total sample were the 904 patients selected at baseline. What percentage of the total do these represent? i.e. what is the total expected sample >70 across these GP practices? Was a power calculation performed? Why exclude those without a medical card? This may limit generalisability as in Ireland those with ill health and poor socioeconomic status (related to risk of adverse healthcare outcomes, frailty, multi-morbidity etc.), are more likely to receive such a card. With government cut backs the free >70s card was means tested and eroded, particularly during the period of study (before the free >70s GP card). This is a limitation introducing sampling bias that should be discussed. Deliberately excluding those with moderate to severe dementia (or at least those presumed to have this based on average MMSE scores at these stages) also results in a select sample, excluding those that often have a high prevalence of ED admissions/readmissions. This is another limitation and should be included in the limitations section of the discussion. Likewise, some patients with an MMSE (which is not routinely performed in GP, even if patients have cognitive impairment) <20 are likely to have the capacity/capability to complete the Pra. The means of following up outcomes is a significant limitation, which isn’t sufficiently stressed in the limitations section. GPs may/may not be aware of all admissions especially those to other hospitals outside of their catchment area or those with a short ‘overnight’ admission boarded in the ED, which would come under the authors definition of an emergency admission. This is a common issue in GP in Ireland, where discharge letters often do not find their way to the GP. Traditionally HIPE data (hospital coding data) is sought from hospitals in a region/individual private hospitals contacted. Poor communication between primary and secondary care in Ireland may also limit the information available to construct the Pra questionnaire. Results The instruments appear to have excellent sensitivity for low risk cases and excellent specificity for high-risk cases at the established cut-off scores. However, these are extreme. A better cut-off should be explored. Perhaps a better balance between sensitivity and specificity could be found using other cut-offs. Youden’s Index or the maximal accuracy approach could be used to calculate these optimal cut-offs for this data rather than depending on cut-offs that were calculated for other samples in different settings. Please include the NPV, PPV, and rate of false positives and negatives in Table 4 and discuss. Also please include a column showing the absolute number (%) of patients having the outcome (i.e. 1 or more admissions in the year). Was the difference in the accuracy (i.e. comparison between the c-

	statistic results) non-significant statistically? I would avoid the term modest, which is used repeatedly to describe the accuracy of the Pra, given that the established interpretation of c-stats scores between 0.6-0.7 is 'poor'. A minor point: ROC curves show 1-specificity on the x-axis. This is not equivocal to sensitivity and specificity. Was reliability tested? Conclusions There is an over-emphasis on the virtues of the high specificity of the instrument. However, the poor sensitivity means that the test is missing many of those at risk. The previous sensitivity of the Pra should also be discussed to balance the conclusions. A balance where sensitivity is very low and specificity is very high should be discussed. What value is that to practitioners? Even though it is not a screen as per se it could be described as a risk-prediction tool to facilitate case-finding. Without sufficient sensitivity many high-risk patients will never get the opportunity to go on for more detailed assessment. Surely this is the main use for the Pra, to case-find person's at high-risk that could then go on to be assessed (as the Pra alone is unlikely to be sufficient to determine management) for further targeted intervention. As said, based on these results the authors' view that this instrument could be used to target patients for inclusion in future trials to tackle ED admissions is overly optimistic and could be toned down. I agree that the modified Pra is slightly more cumbersome and has worse calibration but there is likely to be no significant difference between the two that would make a difference in practice. Indeed, given that the modified Pra doesn't need a questionnaire it is, if anything, more practical. That the Pra is copyrighted and not free to use by institutions is a limitation.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name
Mark Ebell MD, MS
Institution and Country
University of Georgia, United States
Please state any competing interests or state 'None declared':
None declared
Please leave your comments for the authors below
Background Nice overview.
Methods
It isn't clear until Table 1 that the information is being pulled from the GP Record. Or is it? In the text, page 5 line 14 it states "The Pra score was calculated using postal questionnaire data collected in 2010." Yet, Table 1 seems to state that almost all of the data comes from the GP record (which is preferable). Please clarify.
Response
The 'self-rated health' and 'availability of an informal caregiver' items were recorded from the postal questionnaire and the rest of the Pra variables were recorded from the GP medical recorded. I have clarified this in the text as follows: 'The Pra score was calculated using postal questionnaire and GP medical record review data collected in 2010.'
Results
Page 9, lines 33-39 and Figures 1 and 2: The ROC curves do not resemble the typical ROC curve, which is fitted more closely to the actual points. Using this kind of a "stairstep" approach is overly

conservative and underestimates the area under the curve. Also, why only use the 3 pre-specified points? You could use cutpoints of 0.1, 0.2, 0.3, etc to calculate a more accurate curve.

Response

Thank you. We have amended the ROC curve to present cut points in incremental intervals of Pra score 0.1 to the maximum score in this cohort of ≥0.8 for the two versions of the Pra (original and modified). See Figure 1

Table 4: Please report likelihood ratios and predictive values for each stratum, rather than a series of sensitivity/specificity pairs. The whole point of creating risk groups like this is to have richer information about each stratum, which is lost when you simply dichotomize.

Response

Thank you. We have updated Table 4 to include: Cut-point (0, 0.1, 0.2, 0.3 etc. to a maximum to 0.8), Number of patients (n), PPV (%), NPV (%), LR+, LR-. See Table 4.

Obviously, there is a calibration problem, with many fewer patients admitted now that 10-20 years ago. That is not terribly surprising. So why don't you explore how different cutoffs perform, to inform future validation studies? For example, perhaps a moderate risk group of .35 to 0.50, high risk of 0.5 to 0.7, and very high risk of > 0.7? It is OK to do this kind of exploratory analysis if you are clear that the intent is to inform future validation studies.

Response

We have updated Table 5 to include a very high risk group with Pra score of ≥0.7. For the original Pra only 1 participant falls into this group but for the modified Pra 60 participants fall into this group with model specificity of 95%. We have inserted PPVs and NPVs at each risk stratum. We have updated the discussion to comment on this also as follows; 'Based on the findings of the current study, use of the original Pra at cut-point of ≥0.5 may have a role in identifying higher risk patients for enrolment in RCTs in community settings, as it identified a much smaller number of patients as high risk and demonstrated similar predictive accuracy and better calibration than the modified Pra score. However at a higher cut-point of ≥0.7, the modified Pra demonstrated similar predictive accuracy and has the advantage of applicability to the GP record rather than relying on patient questionnaire data so should be explored in future validation studies at this cut-point. The negative predictive value of the Pra is quite high (>80%) at for patients risk-stratified as moderate or high risk and therefore can be useful as a screening test in determining which patients in the community would not benefit from an intervention aiming to reduce future emergency admission rate. However, its positive predictive value is low (approximately 40% for high or very-high risk groups) indicating that a significant proportion of patients who are stratified as high risk will not go on to experience an emergency admission.'

Finally, why not use deciles to create a true calibration curve? This is a post-estimation option in stata, using the HL command with the graph option.

Response

Thank you for this suggestion. We have generated two calibration plots using declines of risk for the original and modified Pra models. See Figures 2 and 3

Discussion

I think you should discuss overall changes in hospital admission rates. Are more patients being managed as outpatients who would have been admitted previously? That would explain the calibration problem. Perhaps you could speculate about an enhanced version of the rule that incorporates the local rates of hospital admission to better calibrate the Pra score for a particular community and healthcare system.

Response

Thank you. We have added the following to the discussion: Case mix variation is a particular issue for external validation studies with heterogeneity of predictor variables and outcomes of interest across study populations.(21) Differences in the prevalence of the outcome of interest, in this case emergency admission, can impact upon the performance of the model. This is one reason for updating the original risk model where the intercept is updated to recalibrate predictive performance for a new setting.(21)

Reviewer: 2
Reviewer Name
Rónán O'Caoimh
Institution and Country
National University of Ireland, Galway
Please state any competing interests or state 'None declared':
None declared
Please leave your comments for the authors below Re bmjopen -2016-012336
Summary
This paper presents an external predictive validation of the Probability of Repeated admission (Pra) risk-prediction tool against a modified version of the same instrument in an Irish primary care setting. The results show that both instruments have poor accuracy in predicting hospitalisation, consistent with similar research with the Pra and other short instruments that measure risk of hospitalisation. I am unconvinced about the utility of the Pra based on these results and disagree with the authors' optimistic view that this instrument could be used to target patients for inclusion in future trials to tackle ED admissions. That it has not been recently externally validated (>10 years as the authors highlight) might relate to the poor accuracy (that the authors themselves found in their own systematic review and was found by Kansagara et al 2011 and O'Caoimh et al 2015. Nevertheless, the study is well conducted and presented.
Abstract
In the conclusion the authors state that the Pra shows modest discrimination. While, this may represent a synthesis of the accuracy and perceived utility of the Pra, the c-statistics presented suggest that the accuracy is poor. The final conclusion should, in my opinion, be toned down to reflect this. The results/findings that are important should be stressed, i.e.
1) The modified version is not significantly more accurate than the original Pra but is less calibrated to predict hospitalisation (which is useful to establish). However, this data could be found from registry data (GP records in this case) making it easier to use than the original in clinical practice.
2) Both have poor accuracy similar to other short-risk prediction instruments used for this purpose.
3) Given the brevity of these instrument and comparability to more detailed and less accessible registry based algorithms, it may be a reasonable choice if a someone wished to use such an instrument.
However, based on the psychometric properties presented here there is no convincing evidence that such an instrument will pay off the time etc. invested to use it, particularly in Irish General Practices, where time is limited.
Response
Thank you for these interesting comments. We have changed the use of the description 'modest' to 'poor' throughout the manuscript. We have included the points regarding clinical utility into the discussion in the abstract, based on conducting additional analysis of the modified Pra at a higher cut-point of ≥0.7 as follows; Future validation studies should examine higher cut-points denoting high risk for the modified Pra, which has practical advantages in terms of application in general practice
Strengths and limitations
This same criticism is applied to the strengths and limitation summary. This is overly positive and does not fully reflect the results. The poor accuracy should be stated. See above points 1-3.
Response
We have added the point regarding terming the performance of the Pra as 'poor' rather than 'modest' throughout the document. In terms of methodological study limitations covered in the discussion we have included issues regarding generalisability as follows; This study recruited patients from 15 urban GP practices in Leinster in the Republic of Ireland and excluded patients with moderate or severe cognitive impairment which may reduce the generalisability of the findings.
Background
Well presented.
A minor point: in the 2nd paragraph a conjunction is missing "of the Pra tool AND identified 10...."

Response
Thank you-this has been corrected.
Methods
Please clarify the sample. Are these practices urban-suburban-rural or a mix? How representative of the total sample were the 904 patients selected at baseline. What percentage of the total do these represent? i.e. what is the total expected sample >70 across these GP practices?
Response
The 15 practices are all in Leinster in the Republic of Ireland and represent a largely urban practice population. A proportionate stratified random sampling approach was used to recruit patients at baseline. There were a total of 4,573 patients aged ≥70 years across the 15 practices. Of these a propionate random sample were selected to participate (n=1,764). A total of 1,487 patients remained eligible following application of exclusion criteria and a total of 904 (response rate=61%) took part in the study at baseline.
We have updated the methods section with the location of the participating practices as follows; This is a two year prospective cohort study of older general practice (GP) patients recruited from 15 practices in Leinster in the Republic of Ireland (2010-2012).
Was a power calculation performed?
Response
A formal power calculation was not performed. However for clinical prediction rule external validation studies it is recommend by the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines that there should be at least 10 primary outcome events (emergency admissions) per variable included in the model. The Pra comprises of 8 variables which necessitates 80 emergency admissions. In this study there were 154 emergency admissions in the prediction period so the study is adequately powered.
Why exclude those without a medical card? This may limit generalisability as in Ireland those with ill health and poor socioeconomic status (related to risk of adverse healthcare outcomes, frailty, multi-morbidity etc.), are more likely to receive such a card. With government cut backs the free >70s card was means tested and eroded, particularly during the period of study (before the free >70s GP card). This is a limitation introducing sampling bias that should be discussed.
Response
It is estimated that 96% of patients aged ≥70 years are in receipt of a medical card even following the introduction of means testing. I have added the following to clarify this: Approximately 96% of all people aged ≥70 years in the Republic of Ireland are in receipt of a GMS card which entitles the holder to free public health services (including GP visits) and prescriptions, subject to a maximum co-payment of €25 monthly (Ref: Health Service Executive (HSE). Primary Care Reimbursement Service statistical analysis of claims and payments 2011. Dublin: HSE, 2013.

Deliberately excluding those with moderate to severe dementia (or at least those presumed to have this based on average MMSE scores at these stages) also results in a select sample, excluding those that often have a high prevalence of ED admissions/readmissions. This is another limitation and should be included in the limitations section of the discussion. Likewise, some patients with an MMSE (which is not routinely performed in GP, even if patients have cognitive impairment) <20 are likely to have the capacity/capability to complete the Pra.
Response
This study is part of a larger study examining the prediction of adverse health outcomes in older people. As such, patients needed the capacity to complete a postal questionnaire as well as a telephone interview regarding adverse drug events. As a result patients with moderate/severe cognitive impairment were excluded at baseline. The MMSE cut-off of <20 was chosen based on the definition of moderate cognitive impairment cited in Folstein MF, Folstein SE, McHugh PR: “Mini-mental state: A practical method for grading the cognitive state of patients for the clinician.” J Psychiatr Res 1975;12:189-198. A documented MMSE in the GP electronic medical record (either in scanned hospital correspondence or administered by the GP) was used to determine this exclusion

criterion.
Of note the original Pra model was not designed for use in a cognitively impaired population as it relies on a patient's ability to recall as part of the questionnaire amongst other items recent healthcare utilisation. The modified Pra may have a role in such a population as it can be applied to the GP medical record.
We have added the exclusion of patients with cognitive impairment as a limitation to the discussion section, with the caveat mentioned, as follows: This study recruited patients from 15 urban GP practices in Leinster in the Republic of Ireland and excluded patients with moderate or severe cognitive impairment which may reduce the generalisability of the findings. However the Pra model was designed as a patient questionnaire which relies on the respondent's ability to recall amongst other items, doctor visits and emergency admission. As such it was not designed for use in a cognitively impaired population. The modified Pra may have a role in this population and this may be tested in future validation studies.

The means of following up outcomes is a significant limitation, which isn't sufficiently stressed in the limitations section. GPs may/may not be aware of all admissions especially those to other hospitals outside of their catchment area or those with a short 'overnight' admission boarded in the ED, which would come under the authors definition of an emergency admission. This is a common issue in GP in Ireland, where discharge letters often do not find their way to the GP. Traditionally HIPE data (hospital coding data) is sought from hospitals in a region/individual private hospitals contacted. Poor communication between primary and secondary care in Ireland may also limit the information available to construct the Pra questionnaire.
Response
All GPs taking part in this study use HealthLink which electronically transfers discharge summaries to the GP from the hospital. In addition all consultations in the GP medical record for the period of interest were reviewed and any additional emergency admissions documented there were included. I'm not clear what data the reviewer is basing his argument that emergency admission discharge summaries often don't reach the GP. The emergency admission rates reported in this study are similar to those reported in this age group in a national Irish Longitudinal Study on Ageing (TILDA), therefore we are confident that our outcome measurement is adequate.
Regarding linkage to Hospital Inpatient Enquiry (HIPE) data there is currently no data linkage agreement in Ireland that allows for linkage of the patient's GMS medical card number to HIPE data. To establish such linkage would have required a process of application to the Data Protection Commissioner to request permissions for linkage which was outside the scope and remit of this study.

Results
The instruments appear to have excellent sensitivity for low risk cases and excellent specificity for high-risk cases at the established cut-off scores. However, these are extreme.
A better cut-off should be explored. Perhaps a better balance between sensitivity and specificity could be found using other cut-offs. Youden's Index or the maximal accuracy approach could be used to calculate these optimal cut-offs for this data rather than depending on cut-offs that were calculated for other samples in different settings.
Response
Thank you. In response to this comment and the other reviewer's comments we have added a new Table 4 presenting PPV, NPV, and likelihood ratios for each incremental increase in the Pra score of 0.1. We have also generated a new ROC curve presenting these risk strata for the original and modified Pra. See Figure 1.
Optimal cut points, identified using both Liu's method and Youden's index, were 0.35 (the threshold for moderate risk) for the original Pra and 0.5 (the threshold for high risk) for the modified Pra.
In addition we have examined a 'very high risk' cohort with Pra cut off of ≥0.7-see Table 5. For external validation studies of risk prediction models it is usually recommended to validate the model as it was developed (as per TRIPOD guidelines). However to inform future validation studies,

considering the poor performance of the Pra in our study, we feel it is reasonable to explore this 'very high risk' risk stratum.

Please include the NPV, PPV, and rate of false positives and negatives in Table 4 and discuss. Also please include a column showing the absolute number (%) of patients having the outcome (i.e. 1 or more admissions in the year).
Response
Thank you. As above Table 4 has been added and Table 5 has been revised to address these comments.

Was the difference in the accuracy (i.e. comparison between the c-statistic results) non-significant statistically?
Response
The difference between the AUC for the original and modified Pra models is not statistically significant (p value=0.38) This has been include in the results section.
I would avoid the term modest, which is used repeatedly to describe the accuracy of the Pra, given that the established interpretation of c-stats scores between 0.6-0.7 is 'poor'.
Response
Thank you-we agree and have changed this terminology throughout the document to 'poor'.

A minor point: ROC curves show 1-specificity on the x-axis. This is not equivocal to sensitivity and specificity.
Response
This has been amended in the main text.

Was reliability tested?
Response
The standard statistical approach in reporting the performance of a risk prediction model in external validation studies is to present model discrimination and calibration (as per the TRIPOD guidelines). Therefore reliability was not tested in this study.

Conclusions
There is an over-emphasis on the virtues of the high specificity of the instrument. However, the poor sensitivity means that the test is missing many of those at risk. The previous sensitivity of the Pra should also be discussed to balance the conclusions. A balance where sensitivity is very low and specificity is very high should be discussed. What value is that to practitioners? Even though it is not a screen as per se it could be described as a risk-prediction tool to facilitate case-finding. Without sufficient sensitivity many high-risk patients will never get the opportunity to go on for more detailed assessment. Surely this is the main use for the Pra, to case-find person's at high-risk that could then go on to be assessed (as the Pra alone is unlikely to be sufficient to determine management) for further targeted intervention.
As said, based on these results the authors' view that this instrument could be used to target patients for inclusion in future trials to tackle ED admissions is overly optimistic and could be toned down. I agree that the modified Pra is slightly more cumbersome and has worse calibration but there is likely to be no significant difference between the two that would make a difference in practice. Indeed, given that the modified Pra doesn't need a questionnaire it is, if anything, more practical. That the Pra is copyrighted and not free to use by institutions is a limitation.
Response
Based on the analysis of the modified Pra at a 'very high risk' stratum of ≥0.7 demonstrating similar predictive accuracy to the original Pra at the high-risk cut-off of ≥0.5, we have updated the discussion to reflect this and to highlight the potential greater clinical utility of the modified Pra which does not need patient questionnaire completion but instead can be applied to the GP medical record as

follows; ‘Based on the findings of the current study, use of the original Pra at cut-point of ≥0.5 may have a role in identifying higher risk patients for enrolment in RCTs in community settings, as it identified a much smaller number of patients as high risk and demonstrated similar predictive accuracy and better calibration than the modified Pra score. However at a higher cut-point of ≥0.7, the modified Pra demonstrated similar predictive accuracy and has the advantage of applicability to the GP record rather than relying on patient questionnaire data so should be explored in future validation studies at this cut-point. The negative predictive value of the Pra is quite high (>80%) at for patients risk-stratified as moderate or high risk and therefore can be useful as a screening test in determining which patients in the community would not benefit from an intervention aiming to reduce future emergency admission rate. However, its positive predictive value is low (approximately 40% for high or very-high risk groups) indicating that a significant proportion of patients who are stratified as high risk will not go on to experience an emergency admission.‘

The conclusion has also been updated as follows; The original Pra tool demonstrated poor discrimination but high specificity in this external validation study and identified a relatively small proportion of patients as high risk. This study suggests that the modified Pra tool, incorporating a multimorbidity measure, is more useful at a higher cut-point indicating high risk which could be examined in future validation studies.

Regarding the Pra model’s low sensitivity and high specificity we would argue that emergency admission is an inherently difficult outcome to predict with the best performing models for this outcome achieving model discrimination of <0.8. (Wallace et al 2015, ref number 5) High sensitivity is particularly important for risk models designed to identify patients for high-stakes diagnoses. For example, the Ottawa ankle rule has model sensitivity of over 99% as clinicians would not use such a tool in clinical practice if there was a risk of missing an ankle fracture. Therefore if the aim is to identify all emergency admissions in a sample then the Pra is not the model of choice. However considering the comparatively poor sensitivity and positive predictive values of other existing models we would argue that a tool, such as the Pra, with high specificity (95%) and negative predictive values of >84% for the moderate and high risk categories, is useful ruling out future admissions in patients classified as moderate or low risk. Therefore intervention efforts can be directed at those stratified as high risk with the understanding that only approximately 1/3 of those identified will actually go on to experience a future emergency admission.

VERSION 2 – REVIEW

REVIEWER	Rónán O'Caoimh National University of Ireland, Galway
REVIEW RETURNED	16-Aug-2016

GENERAL COMMENTS	Thank you for modifying the paper, it is more balanced and reads very well. I suggest adding a line in the methods describing the larger study from which this data were taken so that it is clear that this is a component of a bigger study i.e. to more provide context to the data collection.