February 28, 2017

Dear Dr. Fletcher,

Thank you for the opportunity to revise our paper and respond to the comments of the reviewers and committee members. Please find enclosed our revised manuscript, “Selection of patients for intra-arterial treatment for acute ischaemic stroke: development and validation of a clinical decision tool” (Manuscript ID BMJ.2016.036408), in two separate versions, one with and one without track changes.

We also want to thank the reviewers and committee members for their helpful comments and suggestions. Please find our point-by-point reply and the changes made to the manuscript attached to this letter.

Yours sincerely,

On behalf of all co-authors,

Esmee Venema

Detailed comments from the meeting:

1. First, please revise your paper to respond to all of the comments by the reviewers. Their reports are available at the end of this letter, below.

Response

We thank the editor, the committee members who discussed our manuscript and the reviewers for their thorough review, thoughtful comments and the possibility to respond to the comments. Please find our point-by-point reply and the changes we made to the manuscript below.

2. Please justify your choice of ordinal logistic regression as your model. Did your data meet the assumption of proportionality of the odds ratios?

Response

In the discussion of our statistical analysis plan we explained our choice for the ordinal logistic regression model: “We will use a proportional odds model to analyze the full mRS score as outcome. Formally this model requires the assumption that the odds ratios are the same for each cut-off of the mRS. However previous studies have shown that even if the proportional odds assumption is violated, proportional odds analysis is still more efficient than dichotomization. (McHugh GS, Butcher I, Steyerberg EW, et al. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. Clin Trials 2010;7(1):44-57.) In addition, all recent RCTs on the effect of IAT used the full mRS, analyzed with proportional odds regression as the primary outcome.”

There are several tests available to assess if the proportional odds (PO) assumption is met, but all have been criticized, mainly for being dependent on the sample size: in a small study the PO assumption will always be confirmed because of lack of power while in a large study it will be rejected. Frank Harrell therefore recommends a graphical method for assessing the parallel slopes assumption by plotting the residuals for each cut-off (Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag. 2001). This is shown in the graph below. The partial residuals of the model computed separately for each mRS cut-off, for all predictors in the model. The x-axis shows the mRS cut-offs, the y-axis shows the partial residuals. When these are comparable across all mRS cut-offs, the PO assumption is met. The plot illustrates that for all predictors in our model the residuals for each cut-off are close around the 0 line, so the proportional odds assumption is reasonable to make for all predictors in the model.

However, we believe that dependence of the proportional odds model on the assumption of proportionality is overstressed, as also stated by Senn and Julious (Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians? Stat Med 2009;28:3189-209). If there is consensus on the ordering of the outcome scale (each score on a certain scale is more favourable than a one point lower score), the common odds ratio can be presented and interpreted as a summary estimate of the predictor effect, regardless of violation of the PO assumption. We would prefer not to present the testing of the PO assumption in the paper. However, we are willing to present the testing of the PO assumption if the editorial committee finds it necessary.

Changes to the manuscript

None.

3. Please note in your discussion the limitation of the relatively low discriminatory power of your decision tool. The C statistic is fairly modest, albeit better than what exists already.

Response

We agree that the externally validated ordinal c-statistic was fairly modest (ordinal c-statistic = 0.68). We would like to emphasize however, that the c-statistic for the ordinal outcome is a more conservative measure than the AUC for discrimination between 2 groups with different outcome (for example mRS 0-2 versus mRS 3-6). The ordinal c-statistic assesses discrimination between exact categories of the mRS. The AUC for binary outcome assesses the ordering: in pairs of
patients, is the predicted probability higher for the one with the outcome? Externally validated c-statistics of all different cut-offs were better than the ordinal c-statistic, e.g. 0.72 for good functional outcome and 0.75 for mortality. However, the shift on the ordinal modified Rankin Scale was our primary outcome, and the externally validated c-statistic was lower than the c-statistic in the development sample. Thus, we have to conclude there is inaccuracy of the predictor effects due to overfitting. Nevertheless, the proposed model is currently the best available tool for decision-making in intra-arterial treatment.

Furthermore, when compared to other models commonly used in neurovascular practice (HAS-BLED (c-statistic = 0.65) and CHA2DS2-VASC (c-statistic = 0.61)), our model performs accurately.


Since several comments of both the committee and the reviewers focused on the discussion, we decided to rewrite and reconstruct the second part of the discussion. We constricted the limitations to a single paragraph, and included more explanation on the discriminative ability of the model. Following the limitations, we now discuss possible implications and future directions.

Changes to the manuscript
We replaced the text starting at the fifth paragraph of the discussion section. We discuss model limitations, including the modest c-statistic, applicability of the model, clinical implications and future steps.

We replaced the part from the fifth paragraph of the discussion until the final paragraph with the following:

‘Our study has a number of limitations. The discriminative ability of the model in the external validation was modest. It should be emphasized however, that the c-statistic for the ordinal outcome is a conservative measure. It assesses discrimination between exact categories of the mRS, instead of discrimination between 2 groups with different outcome (for example mRS 0-2 versus mRS 3-6). Externally validated c-statistics of all cut-offs were better than the ordinal c-statistic, e.g. 0.72 for good functional outcome and 0.75 for mortality. Nevertheless, the relatively small sample size and inclusion of interaction terms in the model may have resulted in some optimism and overfitting, despite shrinkage of the regression coefficients. The calibration was also suboptimal; despite the fact that most patients were treated with first generation thrombectomy devices, patients in IMS III had a better outcome than predicted by our model. This could be explained by the stricter patient selection in IMS III (eg. premorbid mRS ≤ 2, age < 82, IV tPA treatment), which resulted in a better prognosis overall. Patients in the IMS III control group had better outcomes than patients in the control group in MR CLEAN (mRS 0-2 = 39% (IMS III with occlusion on CTA) vs 19% (MR CLEAN)), leading to inadequate calibration of our model. Despite its limitations, the currently developed model is the first to predict the effect of IAT for individual patients upon arrival at the emergency department. When compared to other models used in neurovascular practice, HAS-BLED (c-statistic = 0.65) and CHA2DS2-VASC (c-statistic = 0.61), it performs accurately. The predictions made by our decision tool often agree with clinical intuition, which should not be surprising. However, estimates derived from large datasets are preferable to the subjective opinion of a physician, whose experience, no matter how vast, can never match the information contained in large datasets.

Currently some centers withhold IAT in specific subgroups of patients, e.g. low ASPECTS, no collaterals, old age, or M2 occlusion. Indeed, our model predicts no benefit of IAT for some individual patients, especially when a patient has more than one characteristic that negatively affects the effect of IAT. The decision not to treat may be particularly relevant in patients who have to be transferred to an intervention center. The model may help to identify patients without expected benefit of IAT and topple the balance in favor of no treatment. But, perhaps more importantly, our study shows that treatment should not be withheld based on a single characteristic. Some patients belonging to one of the subgroups that are considered as having no benefit of IAT, such as poor collaterals or low ASPECTS, may still benefit from IAT substantially if other characteristics are favourable. This emphasizes the importance of making personalized treatment decisions, instead of using average treatment effects, and shows the need for combining multiple clinical and radiological baseline characteristics instead of withholding treatment based on a single characteristic. However, given that this is the first model for IAT decision making, the predictions of our model should be considered as a starting point for clinical decision making, and not as a final recommendation. Our model was developed in the MR CLEAN database, consisting of a rather unselected population with few selection criteria. Therefore, our model is likely applicable in centers that use few clinical and radiological selection criteria. Future analyses within larger studies may refine the current recommendations and improve the model validity.

Furthermore the second sentence of the conclusion in the abstract and the manuscript:
‘Selection of individual patients for IAT should therefore not be based on single patient characteristics.’

Was replaced with:
‘The model is clinically useful as it aids in distinguishing between individual patients who will experience benefit from intra-arterial treatment for acute ischemic stroke and those who will not.’

4. Please note in your discussion that the calibration for the validation data was over predicting a good outcome. The relative importance of predictions between patients may be accurate but the absolute prognosis was not.

Response
We agree that the predicted absolute prognosis of our model in IMS III was not completely accurate. This can be explained by some important differences between MR CLEAN and IMS III. Differences in selection of patients in IMS III resulted in a patient population with a better prognosis based on observed characteristics, but, given the modest calibration-in-the-large, also on characteristics that were not captured in the model. Despite the use of first generation devices with less relative IAT effect,
outcomes were better. The control group in IMS III had a better outcome on average than the MR CLEAN controls (mRS 0-2 = 39% in IMS III patients with a proven occlusion, vs 19% in MR CLEAN) (Demchuk AM, Goyal M, Yeatts SD, et al. Recanalization and clinical outcome of occlusion sites at baseline CT angiography in the Interventional Management of Stroke III trial. Radiology 2014;273:202-10.). This explains why the overall outcome in IMS III was better than the model predicted, as is shown in figure 2 and in the corresponding calibration-in-the-large of 0.81.

Changes to the manuscript
We emphasize this issue in our discussion, as stated in point 3.
5. Please check the labelling on your figures. We think figure 3 may have been inverted

Response
We have corrected this labelling error.

Comments from Reviewers
Reviewer: 1 (David Barer)
Job Title: Consultant Stroke Physician Institution: City Hospitals Sunderland

The authors have developed and validated a sophisticated model for predicting functional outcome following acute ischaemic stroke, and the likely effect of intra-arterial treatment (IAT) on this outcome. Because it is intended as a decision aid for stroke physicians considering IAT, it incorporates information from CT brain scans and particularly from CT angiography (CTA).

The model incorporates more clinical and imaging information, is derived from and has been validated on much larger datasets than previous similar attempts, and is potentially more useful.

The model was derived using data from the MR CLEAN trial on 500 patients with large intracranial artery occlusions, confirmed by CTA (or MRA), who could receive IAT within 6 hours of stroke onset. The validation dataset comprised 260 patients from the IMS-3 trial, who also had proven arterial occlusions. These represented only 40% of all the 656 patients in IMS-3, since this trial had started before CTA became widely available. Sophisticated methods were used to impute missing data values, adjust for optimism bias due to overfitting of the training dataset, and to allow for possible nonlinear effects. Some of these methods will be unfamiliar to clinicians, so expert statistical review will be important. The text is not easy to follow at times, especially as there seem to be mistakes in the numbering of the references and labelling of some of the figures (see below).

1. Figure 1 is particularly confusing. It is said to illustrate interaction effects between IAT and various prognostic factors and the scales are given as "log ORs" for good functional outcome, yet 4 separate values are shown for IAT+/- vs. Cofactor +/-.

Response
Thank you for this remark, this should indeed be the log odds instead of the log odds ratio, since the effect of the variables on outcome is given per treatment group.

Changes to the manuscript
We replaced the label of the y-axis of Fig 1: 'log OR', with: 'log odds'

2. In the legend of Fig 2, the concept of "calibration in the large" is not explained (how does it differ from the calibration slope?) and I am not sure what is meant by "overall outcome was twice as good as expected". The meaning of the linear bar chart at the bottom of Fig 2 is also not clear to me.

Response
"Calibration-in-the-large" refers to the difference in the mean predicted and mean observed outcome. It indicates if the predicted probabilities are systematically too low or too high. In our study the observed proportion of patients with mRS 0-2 in the validation cohort was higher than predicted by the model (calibration intercept = 0.81, exp(0.81) = OR 2.2)). The calibration slope reflects the strength of the predictors. The bars at the bottom of Figure 2 represents the distribution of the predicted probabilities for patients with (=1) or without (=0) an outcome of mRS 0-2.

We agree that the current legend may be confusing, so we made some changes for a better understanding of the figure. Because it is confusing, we do not use the phrase 'twice as good' anymore.

Changes to the manuscript
We replaced Fig 2 legend:
'Calibration plot for predicted good functional outcome, defined as modified Rankin Scale (mRS) 0-2, in the validation cohort (n=260). Overall, outcome was twice as good as expected (calibration in the large: +0.81, equivalent to an observed to expected ratio of exp(0.81)=2.2), and predictive effects were smaller (calibration slope 0.67 [95% confidence interval (CI) 0.48-0.92]). Discrimination between low and high risk was adequate (c-statistic 0.72 [95% CI 0.65-0.77]).'

With:
'Calibration plot for predicted good functional outcome, defined as modified Rankin Scale (mRS) 0-2, in the validation cohort (n=260). The calibration slope reflects the strength of the predictors. The calibration intercept reflects the calibration-in-the-large, indicating whether predicted probabilities are systematically too low or too high. The overall observed proportion of patients with mRS 0-2 in the validation cohort was higher as to be expected using our model. The linear bar chart shows the distribution of patients with (=1) or without (=0) an observed outcome of mRS 0-2. Discrimination between low and high likelihood of good functional outcome was moderate (c-statistic=0.72 [95% CI:0.65-0.77]).'
3. Labelling errors; Refs 5 is cited out of order in the text and refs 6-10 do not seem to be cited at all. The ref given for the collateral score (Methods, para 1) should surely be 14, not 15. The labelling for Figs 3B and 3C differs between the legend and the figures themselves. There may be other errors in the numbering of references etc, which I have not spotted.

Response

We have updated all references and we have corrected the labelling error of Fig 3. We furthermore checked all other references, added a reference to the latest IAT RCT by Muir et al. (11) and updated ref 37 of Kent et al.

4. General Comments on Study Design; As far as I can tell, the analysis has been done rigorously, using high quality data and following a previously published SAP, but unfortunately there are several factors which could compromise the validity of the model and limit its value as a practical decision aid: The choice of the IMS-3 trial as validation sample, while presumably agreed in advance, was unfortunate since no IAT effect signal was detected in this trial as a whole. IMS-3 was one of the first batch of IAT trials, done largely without the benefit of CTA and using first generation thrombectomy devices rather than stent retrievers, as in MR CLEAN. The better results in the later trials have generally been ascribed to both these factors, yet it appears from Fig 2 that there was a clear benefit for IAT using the older thrombectomy devices in the subsample of IMS used to validate the predictive model. This presumably means that IAT must have been either totally ineffective or harmful in the remaining patients. This unexpected finding requires comment as clinicians may object to the fact that a model developed to predict the benefit of “2nd generation” IAT has been tested in a study of less effective “1st generation” IAT, so that the results in the combined datasets (Fig 3) may not represent the true situation.

Response

The overall treatment effect in IMS III was neutral. The treatment effect in the selection of IMS III patients with proven occlusion was used in our study was lower than 0.92 (95% CI; 1.20-3.08). This means indeed, that in the patients without proven occlusion, the treatment effect was small or absent. This suggests that the better results in the later trials, are at least partly explained by differences in patient selection. As we have selected the patients in whom there was treatment benefit, we still think the external validation in IMS III is valid. However, further validation in more recent RCTs or clinical practice cohorts indeed remains needed.

Changes to the manuscript

Since several comments of both the committee and the reviewers focused on the discussion, we decided to rewrite and reconstruct the second part of the discussion. We discussed this issue in the new limitations section, especially the final part of this paragraph:

“The calibration was also suboptimal; despite the fact that most patients were treated with first generation thrombectomy devices, patients in IMS III had a better outcome than predicted by our model. This could be explained by the stricter patient selection in IMS III (eg. premorbid mRS ≤ 2, age < 82, IV PA treatment),33 which resulted in a better prognosis overall. Patients in the IMS III control group had better outcomes than patients in the control group in MR CLEAN (mRS 0-2 = 39% (IMS III with occlusion on CTA) vs 19% (MR CLEAN)), leading to inadequate calibration of our model.5,33”

5. Although the predicted effect size varies, Figure 3 shows that the model predicts some benefit from IAT in almost every case – including those with low ASPECTS and collateral scores. This makes it difficult to see how much use it would be as a decision aid, since it would be hard to deny a patient invasive treatment on the basis that the expected benefit would be small (health economics considerations are relevant but costs are not discussed in this paper).

Response

For some patients, the model predicts no benefit, or even small harm. We do think that this is relevant information, especially in patients who have to be transferred to an intervention center. The model may help to identify patients without expected benefit of IAT and topple the balance in favor of no treatment.

We acknowledge that this was not clearly stated in our paper and we therefore made some improvements. Furthermore we updated the discussion section and discussed this point and the clinical implications of the model.

Changes to the manuscript

We addressed this in the second paragraph of the Results section – external validation:

“For some patients, who have multiple characteristics that negatively affect treatment benefit, the model predicts no benefit or even harm.”

In the adapted discussion section we discussed clinical applicability and future applications of the model in the last paragraphs:

“Despite its limitations, the currently developed model is the first to predict the effect of IAT for individual patients upon arrival at the emergency department. When compared to other models used in neurovascular practice, HAS-BLED (c-statistic = 0.65) and CHADS2-VASC (c-statistic = 0.61), it performs accurately.34,35 The predictions made by our decision tool often agree with clinical intuition, which should not be surprising. However, estimates derived from large datasets are preferable to the subjective opinion of a physician, whose experience, no matter how vast, can never match the information contained in large datasets.36

Currently some centers withhold IAT in specific subgroups of patients, e.g. low ASPECTS, no collaterals, old age, or M2 occlusion. Indeed, our model predicts no benefit of IAT for some individual patients, especially when a patient has more than one characteristic that negatively affects the effect of IAT. The decision not to treat may be particularly relevant in patients who have to be transferred to an intervention center. The model may help to identify patients without expected benefit of IAT and topple the balance in favor of no treatment. But, perhaps more importantly, our study shows that treatment should not be withheld based on a single characteristic. Some patients belonging to one of the subgroups that are considered as having no benefit of IAT, such as poor collaterals or low ASPECTS, may still benefit from IAT substantially if other characteristics are favourable. This emphasizes the importance of making personalized treatment decisions, instead of using average treatment effects, and shows the need for combining multiple clinical and radiological baseline characteristics instead of withholding treatment based on a single characteristic.37”
However, given that this is the first model for IAT decision making, the predictions of our model should be considered as a starting point for clinical decision making, and not as a final recommendation. Our model was developed in the MR CLEAN database, consisting of a rather unselected population with few selection criteria. Therefore, our model is likely applicable in centers that use few clinical and radiological selection criteria. Future analyses within larger studies may refine the current recommendations and improve the model validity.1

6. According to Fig 1 (if I understand it correctly) there could be a small negative effect of IAT in patients with previous stroke. If so, this could be important, and a plausible biological mechanism should be discussed. In practice, though, the effect largely disappeared after multivariable adjustment and it was not confirmed in the validation sample. These two issues (as well as others, mentioned in the Discussion) cause some misgivings, given that the model is now being developed as a web application for widespread use, but my main concern is that it is extremely difficult to detect harm from treatment using an essentially dichotomous model of this kind.

Response
We defined benefit of IAT indeed dichotomous, as the difference in predicted good outcome (mRS 0-2) after IAT and the predicted good outcome without IAT. This was necessary to calculate predicted and observed proportions for external validation. However our web-application provides the user with bar charts of the mRS in both treatment arms, so every mRS cut-off is shown. Using our app, harm of IAT will be predicted if a patient has all characteristics that negatively influence the effect of treatment (long time to admission, poor collaterals and previous stroke).

Changes to the manuscript
At the end of sentence two in the second paragraph of the Results section – external validation, we added:
‘. . .in the combined derivation and validation cohort’
And at the end of this paragraph we added:
‘For some patients, who have multiple characteristics that negatively affect treatment benefit, the model predicts no benefit or even harm.’
In the final paragraphs of the adapted discussion applicability of the model and predicting harm are now addressed, as stated in the response to reviewer’s fifth point.

7. It is difficult to justify citing the HERMES meta-analysis in support of the statement (Background, para 1) that “there is debate on the selection of candidates for IAT in current practice, because of uncertainty of treatment benefit in specific subgroups”. The HERMES paper emphasizes the consistency in the direction of treatment effects in different patient subgroups.

Response
The results of the HERMES meta-analysis are suggesting treatment effect in different patient subgroups. However, they showed there is still uncertainty in patients with low ASPECTS (0-5), M2 occlusions and low NIHSS. Our multivariable approach shows that some patients with low ASPECTS, low collateral score, low NIHSS at baseline and M2 occlusions still can benefit, if their other characteristics are more favourable.

Changes to the manuscript
We added the reference to the paper of Berkhemer et al. (15) on collateral status and treatment effect since this subgroup was not studied in the cited HERMES paper.

8. Following what has become a standard convention, the authors define good outcome as independent survival (mRS 0-2). Not only is this statistically inefficient, especially for patients with severe strokes (NIHSS>8) as very few can be expected to achieve this level of recovery (see the examples in Fig 4), but it lumps all other levels of functional outcome together (from needing minimal help with some activities to total dependency or death). Thus even a large proportional increase in “good outcomes” may be small in terms of absolute numbers, and could easily be swamped by a small increase in severe vs mild/moderate disability (mRS 5 vs 3-4). Although this does not seem to happen overall, it is important to examine the distribution of mRS outcomes within different levels of stroke severity, so that realistic estimates can be made of both benefits and potential harms of treatment. Failure to do this has been a major weakness of the IPD meta-analysis of IV tPA trials in stroke, where the possible effects of disabling but non-fatal cerebral bleeds have not been taken into account. IC Haemorrhage may be less of a problem with IAT, but other procedural complications need to be considered, including the considerable stress and disappointment involved in transferring patients to tertiary centres, only to find that treatment is impractical or unsuccessful. In effect, a dichotomous analysis which simply classifies mRS 3-6 as “poor outcome” is unethical as it cannot detect treatment effects (positive or negative) at the lower end of the outcome scale and thus denies many patients the opportunity to contribute useful data to the analysis. Ordinal regression analysis uses more of the data but it depends on the validity of the proportional odds assumption, and the authors do not provide any test of this in their data. It is clearly unlikely to hold in trials of IV thrombolysis, where there is an excess of early deaths. By its nature, proportional odds regression cannot detect harms as well as benefits as it assumes an “average” treatment effect at every category boundary. Although the analysis plan for this study was pre-specified, I feel the authors should at least address these issues specifically in the discussion.

Response
We completely agree with the reviewer that it is inefficient and clinically incorrect to use only a dichotomization of mRS for good outcome. We developed our model using ordinal logistic regression which estimates the effect of predictors on the full modified Rankin Scale. The presented odds ratios are common odds ratios for a better outcome on the ordinal modified Rankin Scale. What might have caused confusion is the cut-off that we had to use for the calibration plot. We used the standard mRS 0-2 cut-off and extracted the predictions for this cut-off from the ordinal model. However, when using our app, the
probabilities for each score on the mRS will be displayed, with and without IAT.

In the discussion of our statistical analysis plan we explained our choice for the ordinal logistic regression model: "We will use a proportional odds model to analyze the full mRS score as outcome. Formally this model requires the assumption that the odds ratios are the same for each cut-off of the mRS. However previous studies have shown that even if the proportional odds assumption is violated, proportional odds analysis is still more efficient than dichotomization. (McHugh GS, Butcher I, Steyerberg EW, et al. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. Clin Trials 2010;7(1):44-57.) In addition, all recent RCTs on the effect of IAT used the full mRS, analyzed with proportional odds regression as the primary outcome."

There are several tests available to assess if the proportional odds (PO) assumption is met, but all have been criticized, mainly for being dependent on the sample size: in a small study the PO assumption will always be confirmed because of lack of power while in a large study it will be rejected. Frank Harrell therefore recommends a graphical method for assessing the parallel slopes assumption by plotting the residuals for each cut-off (Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag. 2001). This is shown in the graph below. The partial residuals of the model computed separately for each mRS cut-off, for all predictors in the model. The x-axis shows the mRS cut-offs, the y-axis shows the partial residuals. When these are comparable across all mRS cut-offs, the PO assumption is met. The plot illustrates that for all predictors in our model the residuals for each cut-off are close around the 0 line, so the proportional odds assumption is reasonable to make for all predictors in the model.

However, we believe that dependence of the proportional odds model on the assumption of proportionality is overstressed, as also stated by Senn and Julious (Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians? Stat Med 2009;28:3189-209). If there is consensus on the ordering of the outcome scale (each score on a certain scale is more favourable than a one point lower score), the common odds ratio can be presented and interpreted as a summary estimate of the predictor effect, regardless of violation of the PO assumption. We would prefer not to present the testing of the PO assumption in the paper. However, we are willing to present the testing of the PO assumption if the editorial committee finds it necessary.

Changes to the manuscript

None.

9. As a stroke doctor working in a non-specialist hospital, my main concern is to identify those stroke patients who need immediate CTA, as this will usually involve contacting the regional centre for expert interpretation. Thus the specific question addressed by this paper is of more interest to specialists, as the predictive score only applies to patients in whom angiography has demonstrated occlusion of a major cranial artery. Nevertheless the data could be used to provide useful pointers to who should have a CTA (e.g. "patients aged Y, seen within Z hours" etc). This would make the paper of interest to a more general readership.

Response

It is true that a proven occlusion on CTA is a prerequisite for using our model. There are some clinical prehospital scores to predict the probability of a large vessel occlusion, but these scores are poorly validated and their sensitivity is limited. We therefore developed the model from the assumption that all admitted acute ischemic stroke patients that might be candidate for IAT (acute ischemic stroke, treatable within 6 hours) will undergo noninvasive vessel imaging. In the Netherlands this has now been implemented.

Changes to the manuscript

None.

Reviewer: 2; Guido J. Falcone

Job Title: Assistant Professor, Department of Neurology: Yale School of Medicine

Venema et al report a very interesting study aimed to develop a prediction tool to identify ischemic stroke patients that may benefit from intra-arterial therapy (IAT). They used high-quality data from two randomized controlled trials that have provided key evidence for the current use of IAT in acute cases of ischemic stroke. They identified 11 clinical and radiological variables that associate with functional status. Taken together in a multivariable model, this set of variables provided limited predictive power.

The question tackled by the study is certainly relevant for the field and will no doubt be of interest to the BMJ readership. The authors should be commended for using such high quality data as well as a pre-specified biostatistical plan that addresses all items included in the Tripod checklist.
Addressing the items mentioned below would – in this reviewer’s view – increase the overall impact of the manuscript.

MAJOR POINTS

1. Why were controls (untreated patients) included in the analysis? As these data come from RCTs, randomization guarantees that the treated and untreated groups will be similar except for the intervention (IAT); in this setting, it seems intuitive to pursue this question in exposed (treated) patients only.

Response
Our purpose was to predict treatment benefit for potential IAT candidates, not only outcome. For example, age is a predictor of poor outcome, but its effect does not vary with and without treatment. Thus, age does not affect the expected treatment benefit. Another example is collateral status, which does predict outcome, but mainly in the treatment arm and therefore is a modifier of treatment effect. Such interactions with treatment can only be analysed using data from both the treatment and control arms of a randomised controlled trial.

Changes to the manuscript
None.

2. The overall predictive power of the final model seems limited, as appropriately expressed by the authors when highlighting a c-statistic of 0.68 in the validation cohort. Perhaps the main conclusion should be that baseline clinical and radiological characteristics cannot be used to reliably identify subjects that will benefit from IAT.

Response
We agree that the externally validated ordinal c-statistic was fairly modest (ordinal c-statistic = 0.68). We would like to emphasize however, that the c-statistic for the ordinal outcome is a more conservative measure than the AUC for discrimination between 2 groups with different outcome (for example mRS 0-2 versus mRS 3-6). The ordinal c-statistic assesses discrimination between exact categories of the mRS. The AUC for binary outcome assesses the ordering: in pairs of patients, is the predicted probability higher for the one with the outcome? Externally validated c-statistics of all different cut-offs were better than the ordinal c-statistic, e.g. 0.72 for good functional outcome and 0.75 for mortality. Nevertheless, the shift on the ordinal modified Rankin Scale was our primary outcome, and the externally validated c-statistic was lower than the c-statistic in the development sample. Thus, we have to conclude there is inaccuracy of the predictor effects due to overfitting. Nevertheless, the proposed model is currently the best available tool for decision-making in intra-arterial treatment. Furthermore, when compared to other models commonly used in neurovascular practice (HAS-BLED (c-statistic = 0.65) and CHA2DS2-VASc (c-statistic = 0.61)), our model performs accurately.

2. Lip GY, Nieuwlaat R, Pisters R, et al. Refining clinical risk stratification for predicting stroke and

Since several comments of both the committee and the reviewers focused on the discussion, we decided to rewrite and reconstruct the second part of the discussion. We constricted the limitations to a single paragraph, and included more explanation on the lower discriminative ability of the model.

Following the limitations, we now describe future implications and in a targeted final paragraph of the discussion we discuss possible implications of the presented model.

Changes to the manuscript
We replaced the text starting at the fifth paragraph of the discussion section. We discuss model limitations, including the modest c-statistic, applicability of the model, clinical implications and future steps.

We replaced the part from the fifth paragraph of the discussion until the final paragraph with the following:

‘Our study has a number of limitations. The discriminative ability of the model in the external validation was modest. It should be emphasized however, that the c-statistic for the ordinal outcome is a conservative measure. It assesses discrimination between exact categories of the mRS, instead of discrimination between 2 groups with different outcome (for example mRS 0-2 versus mRS 3-6). Externally validated c-statistics of all cut-offs were better than the ordinal c-statistic, e.g. 0.72 for good functional outcome and 0.75 for mortality. Nevertheless, the relatively small sample size and inclusion of interaction terms in the model may have resulted in some optimism and overfitting, despite shrinkage of the regression coefficients. The calibration was also suboptimal; despite the fact that most patients were treated with first generation thrombectomy devices, patients in IMS III had a better outcome than predicted by our model. This could be explained by the stricter patient selection in IMS III (eg. premorbid mRS ≤ 2, age < 82, IV tPA treatment),33 which resulted in a better prognosis overall. Patients in the IMS III control group had better outcomes than patients in the control group in MR CLEAN (mRS 0-2 = 39% (IMS III with occlusion on CTA) vs 19% (MR CLEAN)), leading to inadequate calibration of our model.5,33

Despite its limitations, the currently developed model is the first to predict the effect of IAT for individual patients upon arrival at the emergency department. When compared to other models used in neurovascular practice, HAS-BLED (c-statistic = 0.65) and CHA2DS2-VASc (c-statistic = 0.61), it performs accurately.34,35 The predictions made by our decision tool often agree with clinical intuition, which should not be surprising. However, estimates derived from large datasets are preferable to the subjective opinion of a physician, whose experience, no matter how vast, can never match the information contained in large datasets.36

Currently some centers withhold IAT in specific subgroups of patients, e.g. low ASPECTS, no collaterals, old age, or M2 occlusion. Indeed, our model predicts no benefit of IAT for some individual patients, especially when a patient has more than one characteristic that negatively affects the effect of IAT. The decision not to treat may be particularly relevant in patients who have to be transferred to an intervention center. The model may help to identify patients without
expected benefit of IAT and topple the balance in favor of no treatment. But, perhaps more importantly, our study shows that treatment should not be withheld based on a single characteristic. Some patients belonging to one of the subgroups that are considered as having no benefit of IAT, such as poor collaterals or low ASPECTS, may still benefit from IAT substantially if other characteristics are favourable. This emphasizes the importance of making personalized treatment decisions, instead of using average treatment effects, and shows the need for combining multiple clinical and radiological baseline characteristics instead of withholding treatment based on a single characteristic.37

However, given that this is the first model for IAT decision making, the predictions of our model should be considered as a starting point for clinical decision making, and not as a final recommendation. Our model was developed in the MR CLEAN database, consisting of a rather unselected population with few selection criteria. Therefore, our model is likely applicable in centers that use few clinical and radiological selection criteria. Future analyses within larger studies may refine the current recommendations and improve the model validity.’

Furthermore the second sentence of the conclusion in the abstract and the manuscript:
‘Selection of individual patients for IAT should therefore not be based on single patient characteristics.’

Was replaced with:
‘The model is clinically useful as it aids in distinguishing between individual patients who will experience benefit from intra-arterial treatment for acute ischemic stroke and those who will not.’

3. The statistical thoroughness is indeed appreciated; at times, however, it becomes difficult to follow the methods and this may become a problem for the BMJ readership, comprised mainly of clinicians. I suggest to streamline the content to highlight intuitive results.

Response
We agree with the reviewer that methods section may be sometimes though to follow. In our statistical analysis plan, our development and validation steps were described in more detail. We aimed to state only the necessary steps in a concise matter in current manuscript, to help the readers understand the results and discussion section.

Changes to the manuscript
At the beginning of the methods we added a paragraph with a concise summary of the performed steps in development and validation of the model:
‘In short we developed a multivariable prediction model in patients in the MR CLEAN trial (Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute Ischemic Stroke in the Netherlands; n=500) and validated this model in a subgroup of patients with an occlusion on CTA included in IMS III trial (Interventional Management of Stroke III, n=260). The primary outcome was the modified Rankin Scale (mRS) at 90 days after stroke. We constructed an ordinal logistic regression model to predict outcome and treatment benefit. This benefit was defined as the difference between the predicted probability of good functional outcome (mRS 0-2) with and without IAT. Variables were selected using uni- and multivariable selection steps (P<0.15).’

MINOR POINTS

ABSTRACT
1. You may want to rephrase the statement “adequate discriminative ability” in light of the c-statistics values presented in the abstract.

Response
We agree that the externally validated c-statistic was 0.68 for the ordinal model and 0.72 for the prediction of good functional outcome is suboptimal (as discussed in your second comment).

Changes to the manuscript
In the abstract “adequate discriminative ability” was changed into ‘moderate discriminative ability’.

INTRODUCTION

2. In paragraph 1, before mentioning the NNT, please include the effect estimate for IAT.

Response
We have included the odds ratio.

Changes to the manuscript
In the introduction ‘with a number needed to treat of 5’ was replaced by ‘with a number needed to treat of 5, (odds ratio (OR) = 2.35; 95% confidence interval (CI) 0.85–2.98)’.

3. Please remove the clinical examples from paragraph 2. Instead, I suggest to summarize existing studies looking at this same question.

Response
We chose to use the clinical examples to make our paper more appealing for the BMJ readership. We have summarized existing studies in paragraph 2 of the discussion. Of course we are willing to change this if the editorial committee prefers this.

Changes to the manuscript
None.

4. I would also mention that the uncertainty about benefit when using IAT extends to not only some RCT subgroups but also populations not included in these trials.

Response
Indeed, we added this comment to the introduction.

Changes to the manuscript
In the first paragraph of the background section, we added to the last sentence: ‘... because of uncertainty of treatment benefit in specific subgroups.’

METHODS

5. Please define NCCT when first using the acronym.

Changes to the manuscript
We replaced ‘NCCT’ with ‘non-contrast computed tomography (NCCT)’ in the second paragraph of the methods section.

6. Please define mRS when first using the acronym in the main body of the manuscript (it’s defined in the abstract only).

Changes to the manuscript
We replaced ‘mRS’ with ‘modified Rankin Scale (mRS)’ in the first paragraph of the Methods section.

7. “Proportional odds regression modeling” could be replaced by ordinal logistic regression, which assumes proportional odds (but this assumption may or may not hold).

Changes to the manuscript
We replaced ‘proportional odds regression modeling’ with ‘ordinal logistic regression’ throughout the manuscript. In the Methods section we added ‘ordinal logistic regression, which assumes proportional odds’.

8. What model did you fit in the secondary analysis that used a dichotomized mRS? Logistic regression?

Response
We did not fit a different model but we extracted the predicted probabilities for each category of the mRS (including 0+1+2) from the ordinal logistic model. This is described in the first paragraph of the ‘model development’ section in the Methods.

Changes to the manuscript
None.

9. How was the p-value threshold of 0.15 selected to keep covariates in the model?

Response
In prediction research it is advisable to use a more lenient p-value threshold for selection of predictors to prevent overfitting of the model, instead of a p-value < 0.05. Commonly used threshold vary from 0.1 to 0.2 (Royston Patrick, Moons Karel G M, Altman Douglas G, Vergouwe Yvonne. Prognosis and prognostic research: Developing a prognostic model BMJ 2009; 338 :b604), and we chose the value of 0.15.

Changes to the manuscript
None.

10. Was collinearity factored in in any way when building the model?

Response
We assessed collinearity based on clinical relevance and checked for potential collinearity during the selection of predictors.

Changes to the manuscript
None.
11. The collateral score is a 3-point score, it’s probably better to say that a test for trend across these 3 categories was used.

Response
We used the collateral score as a 4 point score, ranging from 0-3. (Tan et al. CT angiography clot burden score and collateral score. 2009 Am J of neuroradiology) We have performed a trend test and found that it was reasonable to analyse the collateral score continuously.

Changes to the manuscript
None.

RESULTS

12. When presenting statistically significant predictors of outcome in paragraph 2, please include the p-value to give the reader an idea of the strength of the association without having to look at the table.

Response
We have included the p-values of the predictors in the Results section.

Changes to the manuscript
In the second paragraph of the Results section, we have replaced ‘The strongest predictors in multivariable analysis were age, baseline NIHSS, systolic blood pressure, history of ischaemic stroke, diabetes mellitus, pre-stroke mRS, ASPECTS, location of occlusion and collateral score.’ With ‘The strongest predictors in multivariable analysis were age (p<0.01), baseline NIHSS (p<0.01), systolic blood pressure (p<0.01), history of ischaemic stroke (p=0.03), diabetes mellitus (p=0.02), pre-stroke mRS (p<0.01), ASPECTS (p<0.01), location of occlusion (p=0.03) and collateral score (p<0.01).’

13. "Interactions with relative treatment effect were found in univariable analysis for history of ischaemic stroke, atrial fibrillation, collateral score and time to groin puncture (all p<0.15). The effect of IAT was non proportional in relation to these variables." The second phrase seems a bit redundant: if the interaction test was significant, it means that the association for one variable was not the same across strata of the second variable including in the product term (or vice versa).

Response
This is indeed an unnecessary sentence and it was removed.

Changes to the manuscript
We deleted the sentence.

14. Congratulations to the authors for the on-line prediction tool: it’s simple and elegant. I wonder, though, if this is the right time to deploy it given limited predictive ability of the model.

Response
In response to the reviewers third major comment we updated our discussion section. The limited
predictive ability was discussed at the limitations paragraph and the last paragraphs discuss the model applicability.

Changes to the manuscript
As stated in the response to the reviewer’s second major comment.

DISCUSSION

15. I suggest including a paragraph explaining what are the likely next steps to more effectively tackle patient selection in IAT.

Response
Thank you for this suggestion.
Changes to the manuscript
We have rewritten the discussion and focused a total paragraph on applicability and future steps, as stated in the response at reviewer’s second major comment.

16. Please abbreviate and consolidate the limitations in a single paragraph.

Response
We have rewritten our discussion and consolidated the limitations in a single paragraph.

Changes to the manuscript
As stated in the response at reviewer’s second major comment.

TABLES

17. For table 2, please state that the effect estimates express the odds of a better outcome.

Changes to the manuscript
‘Table 2. Main effects in the derivation cohort (n=500).’

is replaced by;

‘Table 2. Main effects in the derivation cohort (n=500). Presented common odds ratios reflect the effect on the reversed modified Rankin Scale (odds ratio >1 corresponds with better functional outcome).’

SUPPLEMENT

18. If accepted for publication, I would include the revised version of the protocol (as opposed to the raw version showing the track changes).

Response
The study protocol is accepted for publication in BMJ Open. We will add the final version of the protocol to this manuscript.