PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form 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 risk prediction models for post-stroke mortality in Berlin

Authors

Reitzle, Lukas; Rohmann, Jessica L; Kurth, Tobias; Audebert, Heinrich; Piccininni, Marco

VERSION 1 - REVIEW

Reviewer	**1**
Name	**Kumar, Amit**
Affiliation	**University of Utah Health, Physical Therapy**
Date	**13-Sep-2024**
COI	**None**

The significance of the paper cannot be overstated. Nonetheless, offering more detailed information in the introduction and making substantial improvements in the methods would greatly enhance its value. Below are some suggested recommendations for the authors to consider during the review process.

1. Introduction: Past work on the Bray and Smith models is of utmost importance for this research. The introduction lacks significance for this work and needs more evidence of past work. Simply citing one systematic review from Fahey may not do justice to the significance of these models.

2. NIHSS is a strong predictor of mortality. What extra do these two models add to predicting mortality? PMCID: PMC9411458 DOI: 10.1007/s11606-021-07162-0

3. Method: It would greatly benefit the readers of this journal if the author could provide clear definitions of the various variables in both the Bray model and the Smith model.

4. Outcome: In the paper, the author provides two different definitions of mortality for both models. Comparing two different models is difficult if one predicts 30-day all-cause mortality and another predicts in-hospital mortality. It would be beneficial to compare both models to predict the same outcomes (30-day all-cause mortality and in-hospital mortality). This comparison would provide a clearer picture of each model's strengths and limitations.

5. Please clarify what other variables have been added to the prediction model. If the prediction model includes age, gender, and NIHSS, please explain the impact on the c-statistics. In this case, start with a base model that includes demographics and then add variables from the Smith or Bray model.

Example

Base Model: C stat

Base Model + Bray model: C stat

Base Model + Smith model: C-stat

Check these paper PMCID: PMC5097024 and PMCID: PMC5233915 DOI: 10.1093/gerona/glw148

https://jamanetwork.com/journals/jamasurgery/fullarticle/2790270

6. Tables 1 and 2. It would be beneficial to include information on clinical characteristics (for example, hospital length of stay, ICU stay, patient received TPA, comorbidities, and amount of rehab) during the acute stay. These factors may impact in-hospital and 30-day mortality.

Reviewer	**2**
Name	**Kelson, Zoe**
Affiliation	**University of Exeter, Mathematics**
Date	**26-Feb-2025**
COI	**None**

This study aimed to assess the performance of two prediction models for post-stroke mortality in Berlin, Germany.

Reviewer comments:

"data from the Berlin-SPecific Acute Treatment in Ischemic or hAemorrhagic stroke with Long term follow–up (B–SPATIAL) registry."

and

"Adult patients, admitted within 6 hours after symptom onset and with an ICD-10 discharge diagnosis of ischemic stroke, hemorrhagic stroke or transient ischemic attack at one of 15 hospitals with stroke units between January 1st, 2016 and January 31st, 2021."

and

"For the validation of Bray et al.’s model, we included 7,879 patients (mean age 75; 55.0% men)"

and

"For Smith et al.'s model, we performed the validation among 1,931 patients (mean age 75; 56.2% men)"

and

"The strengths of this study include the prospective, multicenter design of the B-SPATIAL registry with coverage of all 15 Berlin stroke units during a 5-year period. Therefore, the B-SPATIAL registry for adult stroke patients can be considered representative for the population of stroke patients in Berlin and comprises detailed information on demographics and clinical characteristics, with low loss to follow-up, especially for mortality endpoints"

The authors confirm the included cohort can be considered to be representative.

Can the authors please consider exploring differences by hospital?

"For both models, we excluded patients who were transferred from a hospital not participating in the B-SPATIAL registry or with missing values for one of the predictors or the outcome. "

Can the authors please clarify how much data was missing, and whether it can be considered to be missing at random?

"We assessed the discriminatory ability of the two prediction models by calculating the concordance statistic (c-statistic) and visualizing the receiver operating characteristics (ROC)-curve. We computed 95% confidence intervals for the c-statistic based on 2,000 stratified bootstrap replicates"

The authors have suitably assessed model performance metrics.

"we reran our analysis excluding patients with unknown or missing mode of arrival."

and

"we investigated the performance of both models when classifying all patients with final diagnosis of TIA as ischemic stroke patients"

The use of sensitivity analyses help to demonstrate the robustness of the study findings.

Can the authors please consider undertaking subgroup analyses, by gender or age group for example?

Did the authors consider re-calibrating the scores for the cohort in hand, to explore any possible improvement in performance for either model?

"The prediction model for in-hospital mortality after stroke by Smith et al. could only be validated with data from 3 of the 15 hospitals, routinely collecting data on all relevant predictors." [strengths and limitations of this study]

and

"Some limitations should be considered when interpreting our results. As Berlin is a densely populated city with several stroke units, the availability of stroke care might be different compared to other regional settings in Germany and Central Europe. Therefore, our results may not generalize to different settings, such as rural areas. Furthermore, the B-SPATIAL registry only contains information on patients with hospital arrival within 6 hours of symptom onset, since this was the time window for reperfusion treatment eligibility when the registry commenced. However, a substantial proportion of stroke patients present to hospitals later than 6 hours after onset,15, 16 and the performance of these prediction models might differ for these patients.

Only three of 15 registry hospitals routinely documented history of hyperlipidemia, coronary artery disease and history of stroke or TIA. Therefore, we could only validate Smith et al.'s model in this subsample, which composed 30% of the full validation sample. Furthermore, as overall in-hospital mortality risk was low in our setting, only 105 in-hospital deaths were observed in this subsample, which decreased the power of the analysis and somewhat limits the interpretation of the calibration plot due to higher uncertainty, especially in the lower decile groups, in which only few events were observed. For both models, we excluded patients with missing information on at least one of the predictors (except mode of arrival), as this information was missing only for a small number of stroke patients. "

A discussion on the study limitations has been provided.

Thanks for providing a copy of the TRIPOD checklist.

VERSION 1 - AUTHOR RESPONSE

Reviewer 1

The significance of the paper cannot be overstated. Nonetheless, offering more detailed information in the introduction and making substantial improvements in the methods would greatly enhance its value. Below are some suggested recommendations for the authors to consider during the review process.

>> Response: We appreciate the reviewer's comment about the significance of our work, and we have made an effort to improve our reporting in the Introduction and Methods sections in accordance with the reviewer's suggestions to improve accessibility for readers. Please find our detailed responses below. <<

1. Introduction: Past work on the Bray and Smith models is of utmost importance for this research. The introduction lacks significance for this work and needs more evidence of past work. Simply citing one systematic review from Fahey may not do justice to the significance of these models.

>> Response: We appreciate the opportunity to clarify here. Having data from Berlin stroke patients in the B-SPATIAL registry, we explored the possibility of validating existing prediction models in our setting, which is an important stage of the life cycle of clinical risk prediction models (see e.g., https://doi.org/10.1186/1471-2288-14-40, https://doi.org/10.1136/heartjnl-2011-301247). Our choice of the Bray et al. and Smith et al. models hinged on the fact that the relevant predictor variables were available in our dataset. Furthermore, according to a literature search, which we ran again while preparing this revision to check for any new publications, neither the Bray et al. nor Smith et al. models have yet been subjected to external validation in Germany. This is the knowledge gap our paper aims to address.
We agree with the reviewer that the way we presented the models may make it seem that these are the only two models for post-stroke mortality. Therefore, we have now added multiple references to other prediction models in the Introduction and also point readers toward the relevance of stroke severity as a predictor, citing some of the work suggested by the reviewer in the second comment:

"Despite the abundance of existing prediction models for post-stroke outcomes, only a small fraction have been externally validated.2 Among the most frequently used predictors were demographic characteristics (e.g., age and sex), stroke severity as measured by the National Institutes of Health Stroke Scale (NIHSS), stroke type, and comorbidities.2 The variable NIHSS, alone, has shown high predictive performance for early mortality after acute stroke5 and is often used in prediction models for post-stroke mortality.6, 7
Two prediction models for post-stroke mortality including the NIHSS and other routinely collected variables were developed by Bray et al. (2014) using data from the Sentinel Stroke National Audit Program (SSNAP) in the United Kingdom3 and by Smith et al. (2010) using data from the Get With the Guidelines (GWTG) Stroke Program in the United States.8" <<

2. NIHSS is a strong predictor of mortality. What extra do these two models add to predicting mortality?
PMCID: PMC9411458 DOI: 10.1007/s11606-021-07162-0

>> Response: We agree with the reviewer that the stroke severity, as measured by the National Institutes of Health Stroke Scale (NIHSS), is, by itself, a strong predictor of post-stroke mortality and that this information might be useful to readers. As mentioned in our reply to the previous point, we have now added the following statements with some references suggested by the reviewer to the Introduction:
"Among the most frequently used predictors were demographic characteristics (e.g., age and sex), stroke severity as measured by the National Institutes of Health Stroke Scale (NIHSS), stroke type, and comorbidities.2 The variable NIHSS, alone, has shown high predictive performance for early mortality after acute stroke5 and is often used in prediction models for post-stroke mortality.6, 7"

Furthermore, to understand how well the models' performances compare to NIHSS alone, we have added an analysis assessing the discriminatory ability (c-statistic) of NIHSS for both outcomes (30-day and in-hospital mortality) in our validation datasets. We have updated the Methods accordingly and have modified the Results as follows:
For Bray et al.'s model:
"The c-statistic for 30-day mortality, the model's intended outcome, was 0.865 (95%CI: 0.851-0.879). For comparison, the NIHSS alone showed a c-statistic of 0.838 (95%CI: 0.823-0.853)."
For Smith et al.'s model:
"The corresponding c-statistic was 0.891 (95%CI: 0.864-0.918). For comparison, the c-statistic for in-hospital mortality of NIHSS alone was 0.868 (95%CI: 0.833-0.903)."

Indeed, as seen in other settings, the discrimination ability of the NIHSS alone for post-stroke mortality is quite high. We now reflect on this finding also in light of prior work, adding the following sentence to our Discussion with the suggested reference:
"Our results underscore the high predictive ability of NIHSS, which as a single predictor attained a c-statistic of more than 0.83 for both mortality outcomes, comparable to previous studies.7, 16" <<

3. Method: It would greatly benefit the readers of this journal if the author could provide clear definitions of the various variables in both the Bray model and the Smith model.

>> Response: We agree with the reviewer that a clear definition of predictors and outcomes is important in external validation studies. As both model development studies used registry data, the data were entered by clinicians following the instructions of the respective registry's data collection mask. For Bray et al.'s model, we have now added the following additional details to our Methods section:
"In the original development study, all variables were directly entered in a secure web portal by clinical teams in accordance with the SSNAP registry.3"

Similarly, for Smith et al.'s model, we added:
"In the GWTG registry, used in the development study, clinicians used an internet-based tool for data entry.8"

Unfortunately, we could not find more specific information about the predictors included in the original Bray et al. and Smith et al. models. We tried, however, to specify our predictor and outcome definitions as clearly as possible (see new changes in the Methods). <<

4. Outcome: In the paper, the author provides two different definitions of mortality for both models. Comparing two different models is difficult if one predicts 30-day all-cause mortality and another predicts in-hospital mortality. It would be beneficial to compare both models to predict the same outcomes (30-day all-cause mortality and in-hospital mortality). This comparison would provide a clearer picture of each model's strengths and limitations.

>> Response: Thank you for raising this point. The original models were built using two different mortality outcomes (30-day mortality and in-hospital mortality), so in our primary validation, we used the outcomes the models were designed to predict. However, we appreciate the reviewer's suggestion to assess each model's performance in predicting both mortality outcomes since data for both outcomes were available in the B-SPATIAL registry. We have included a series of new analyses in which each model predicts the outcome the other model was designed to predict, reporting the respective c-statistics.
As the reviewer is likely aware, the head-to-head comparison is only meaningful for discrimination performance, while comparing the calibration performance of the models for the different outcomes directly is not possible (as the probability of dying within 30 days after stroke and in-hospital mortality are quite different).

Accordingly, we have now described the additional analyses in the Methods:
"In addition, we assessed the discriminatory ability of the NIHSS alone for both outcomes and computed the c-statistic for Bray et al.'s model predicting in-hospital mortality and Smith et al.'s model predicting 30-day mortality."

We report these results as follows in the Results:
"When instead using Bray et al.'s model to predict in-hospital mortality in this validation dataset, we obtained a c-statistic of 0.873 (95%CI: 0.858-0.888)."
and
"When instead using Smith et al.'s model to predict 30-day mortality in this validation dataset, the c-statistic was 0.873 (95%CI: 0.847-0.899)."<<

5. Please clarify what other variables have been added to the prediction model. If the prediction model includes age, gender, and NIHSS, please explain the impact on the c-statistics. In this case, start with a base model that includes demographics and then add variables from the Smith or Bray model.
Example
Base Model: C stat
Base Model + Bray model: C stat
Base Model + Smith model: C-stat
Check these paper PMCID: PMC5097024 and PMCID: PMC5233915 DOI: 10.1093/gerona/glw148
https://jamanetwork.com/journals/jamasurgery/fullarticle/2790270

>>Response: While we appreciate the reviewer's suggestion to fit new prediction models for post-stroke mortality, we reiterate that our purpose was not to develop a new model, but to subject existing models in their original form to external validation in our setting. External validation does not involve refitting the models (see e.g., https://doi.org/10.1186/1471-2288-14-40, https://doi.org/10.1136/heartjnl-2011-301247). Therefore, we did not add any new variables or subtract existing variables. We realize that we may not have conveyed this aim clearly enough in the Introduction, and we have now rephrased the final paragraph of the Introduction as follows:

"Two prediction models for post-stroke mortality including the NIHSS and other routinely collected variables were developed by Bray et al. (2014) using data from the Sentinel Stroke National Audit Program (SSNAP) in the United Kingdom3 and by Smith et al. (2010) using data from the Get With the Guidelines (GWTG) Stroke Program in the United States.8 Though these models have already been subjected to validation studies in their respective originating countries9, to date, both models have only undergone external validation in the China National Stroke Registry.10-12 Our aim was to conduct an external validation4 study to assess calibration and discrimination performances of Bray et al. (2014)3 and Smith et al. (2010)8 prediction models for post-stroke mortality among Berlin stroke patients."

In our external validation, we followed the suggestions in Moons et al. (https://doi.org/10.1136/heartjnl-2011-301247). After preparation of the dataset complete with predictor variables similar to the original studies, we assessed each model's predictive performance, i.e., calibration and discrimination. The predictions were obtained using the models in their original form (i.e., same regression coefficients as the published formulas). Therefore, we did not add or remove any variables when performing the external validations since this would alter the meaning of the original coefficients (from the development studies). However, it is possible to compare the discrimination performance of the overall model with the discrimination performance of only one predictor (obtaining these c-statistics does not require regression coefficients). As detailed in the prior responses, we have now added a comparison with NIHSS alone (which could be considered a "base model") and hope this addresses the reviewers' points.<<

6. Tables 1 and 2. It would be beneficial to include information on clinical characteristics (for example, hospital length of stay, ICU stay, patient received TPA, comorbidities, and amount of rehab) during the

acute stay. These factors may impact in-hospital and 30-day mortality.

>> Response: We agree with the reviewer's suggestion and have now added further details to characterize the study population. Specifically, for the validation of Bray et al.'s model, we added the length of hospital stay, whether patients received systemic thrombolysis, and information about the comorbidities diabetes mellitus and hypertension (see Table 1). For the validation of Smith et al.'s model, we added information about hypertension, systemic thrombolysis, and the length of hospitalization (Table 2). Information on rehab and ICU stay was unfortunately not available in the B-SPATIAL registry dataset.<<

Reviewer 2

This study aimed to assess the performance of two prediction models for post-stroke mortality in Berlin, Germany.

Reviewer comments:

"data from the Berlin-SPecific Acute Treatment in Ischemic or hAemorrhagic stroke with Long term follow–up (B–SPATIAL) registry."
and
"Adult patients, admitted within 6 hours after symptom onset and with an ICD-10 discharge diagnosis of ischemic stroke, hemorrhagic stroke or transient ischemic attack at one of 15 hospitals with stroke units between January 1st, 2016 and January 31st, 2021."
and
"For the validation of Bray et al.'s model, we included 7,879 patients (mean age 75; 55.0% men)"
and
"For Smith et al.'s model, we performed the validation among 1,931 patients (mean age 75; 56.2% men)"
and
"The strengths of this study include the prospective, multicenter design of the B-SPATIAL registry with coverage of all 15 Berlin stroke units during a 5-year period. Therefore, the B-SPATIAL registry for adult stroke patients can be considered representative for the population of stroke patients in Berlin and comprises detailed information on demographics and clinical characteristics, with low loss to follow-up, especially for mortality endpoints"
The authors confirm the included cohort can be considered to be representative.

>> Response: We thank the reviewer for highlighting the key aspects of our study. As the B-SPATIAL registry collected data from adult stroke patients presenting to all 15 hospitals with stroke units in Berlin during the study period, we believe the data can be considered representative for the Berlin population of stroke patients.<<

Can the authors please consider exploring differences by hospital?

>> Response: Our primary aim was to evaluate the performance of the prediction models for post-stroke mortality among Berlin stroke patients. For this purpose, we used the B-SPATIAL registry, which consisted of patient characteristics and process parameter data for stroke patients presenting to the 15 hospitals with stroke units across Berlin during the study period. The main purpose of this registry was quality assurance, and the individual hospitals agreed to contribute their data under the condition that analyses prevent identification of the individual respective hospitals. Therefore, it is not possible for us to compare mortality between individual hospitals.<<

"For both models, we excluded patients who were transferred from a hospital not participating in the B-SPATIAL registry or with missing values for one of the predictors or the outcome. "
Can the authors please clarify how much data was missing, and whether it can be considered to be missing at random?

>> Response: While the number of missing values for the main variables is quite low, we agree that missing values might influence the results of this external validation study. In the flow chart (Figure 1), we have specified the number of missings by predictor for each model. Furthermore, in addition to the complete case analysis, we have now added an additional sensitivity analysis imputing missing values using multiple imputation. We have described this in detail in the Methods as follows:
"Finally, we assessed calibration and discrimination of both models after imputing the predictors' missing values by Multiple Imputation by Chained Equations. Specifically, for each model's validation, we imputed 5 datasets using only the model-specific predictors and outcome in the imputation."

Overall, the results of the analyses using the imputed datasets showed similar model performance for both Bray et al.'s and Smith et al.'s models compared to the main analysis. We have added these findings to the Results (and Supplemental Material) as follows:
For Bray et al.'s model: "In a second sensitivity analysis, in which we used multiple imputation, a total of 8,366 stroke patients were included, of whom 951 (11.4%) died within 30 days. The observed mortality was higher compared to the main analysis, but the conclusions did not fundamentally change. The model underestimated 30-day mortality in the highest risk individuals (to a slightly greater extent than as was assessed in the main analysis; Figure S4). The model's calibration intercept was 0.52 [95%CI: 0.37-0.67]), and the calibration slope was 1.12 [95%CI: 1.04-1.19]). The c-statistic obtained after multiple imputation was 0.870 [95%CI: 0.855-0.884], similar to the main analysis."
For Smith et al.'s model: "In a further sensitivity analysis, in which we used multiple imputation, a total of 2,052 ischemic stroke patients were included, of whom 117 (5.7%) died during the hospital stay. After imputation, the model's calibration intercept (1.26 [95%CI: 0.70-1. 81]), slope (1.44 [95%CI: 1.22-1.66]), and calibration plot (Figure S9) as well as c-statistic (0.893 [95%CI: 0.864-0.917]) were similar to the main analysis."
These findings suggest that the exclusion of missing values in the main analysis did not meaningfully change the interpretation of the transportability of both models to the Berlin setting. We added two sentences (one for each model) in the Discussion reflecting on this point:
"Overall, our conclusions did not differ after multiple imputation or when stratifying by sex."
"The findings of the external validation did not differ substantially after multiple imputation."<<

"We assessed the discriminatory ability of the two prediction models by calculating the concordance statistic (c-statistic) and visualizing the receiver operating characteristics (ROC)-curve. We computed 95% confidence intervals for the c-statistic based on 2,000 stratified bootstrap replicates"
The authors have suitably assessed model performance metrics.
"we reran our analysis excluding patients with unknown or missing mode of arrival."
and
"we investigated the performance of both models when classifying all patients with final diagnosis of TIA as ischemic stroke patients"
The use of sensitivity analyses help to demonstrate the robustness of the study findings.

>> Response: Thank you for these valuable comments. Similar to other studies, we tested the robustness of different assumptions made in the paper e.g., with regard to handling of missing information or classification of stroke type. We report several sensitivity analyses in the manuscript (some additional ones added based on the reviewers' suggestions) and include supplemental figures in the Supplement). These are detailed in the Methods section and reported in the Results.<<

Can the authors please consider undertaking subgroup analyses, by gender or age group for example?

>> Response: We agree that it may be interesting to explore whether the models' performances in our setting might differ for different subgroups. We have now added a subgroup analysis by sex and report the results in the main text for Bray et al.'s model as follows:
"In the subgroup analysis by sex, Bray et al.'s model showed similar performance for male and non-male patients with regard to calibration (Figure S1) and discrimination (c-statistic of 0.858 [95%CI: 0.836-0.880] for males and 0.865 [95%CI: 0.847-0.883] for non-males; Figure S2)."
For Smith et al.'s model:
"Compared to non-male patients, the calibration of Smith et al.'s model seemed slightly better among male patients, as the underestimation of the predicted risk was lower in the highest risk decile groups (Figure S5). Discrimination ability seemed higher for male patients with a c-statistic of 0.914 (95%CI: 0.881-0.946) compared to non-male patients (c-statistic: 0.867 [95%CI: 0.825-0.908]) (Figure S6)."
Please see also the corresponding figures (calibration plots and ROC curves) in the Supplemental Materials.
We added two sentences (one for each model) in the Discussion commenting on these results:
For Bray et al.'s model "Overall, our conclusions did not differ after multiple imputation or when stratifying by sex."
For Smith et al.'s model "We observed slightly better model performance for male stroke patients."<<

Did the authors consider re-calibrating the scores for the cohort in hand, to explore any possible improvement in performance for either model?

>> Response: We appreciate the suggestion to consider recalibrating the models. The purpose of our study was not to update the model but to externally validate the prediction models of Bray et al. and Smith et al. in Berlin stroke patients. In doing so, we followed the suggestions in Moons et al. (https://doi.org/10.1136/heartjnl-2011-301247). As we mentioned in our response to the first reviewer,

we realize we may not have made this aim explicit enough and have now rephrased the final sentence of the Introduction as follows to clarify:
"Our aim was to conduct an external validation4 study to assess calibration and discrimination performances of Bray et al. (2014)3 and Smith et al. (2010)8 prediction models for post-stroke mortality among Berlin stroke patients."

However, to assess calibration, we relied on the "logistic recalibration framework," which involves evaluating the "closeness" of observed mortality and risk prediction by fitting a logistic regression with the observed outcome as the dependent variable and the logit of the predictions as the independent variable. For each validation, we report the value of the intercept (calibration intercept) and slope (calibration slope) of these logistic regressions. This means that interested readers could use our calibration coefficients to better "adjust" the predictions of the models to the Berlin setting. Indeed, the recalibration framework is considered a simple and efficient way to recalibrate prediction models (see, e.g., https://doi.org/10.1136/heartjnl-2011-301247, https://link.springer.com/book/10.1007/978-3-030-16399-0). <<

"The prediction model for in-hospital mortality after stroke by Smith et al. could only be validated with data from 3 of the 15 hospitals, routinely collecting data on all relevant predictors." [strengths and limitations of this study]
and
"Some limitations should be considered when interpreting our results. As Berlin is a densely populated city with several stroke units, the availability of stroke care might be different compared to other regional settings in Germany and Central Europe. Therefore, our results may not generalize to different settings, such as rural areas. Furthermore, the B-SPATIAL registry only contains information on patients with hospital arrival within 6 hours of symptom onset, since this was the time window for reperfusion treatment eligibility when the registry commenced. However, a substantial proportion of stroke patients present to hospitals later than 6 hours after onset,15, 16 and the performance of these prediction models might differ for these patients.
Only three of 15 registry hospitals routinely documented history of hyperlipidemia, coronary artery disease and history of stroke or TIA. Therefore, we could only validate Smith et al.'s model in this subsample, which composed 30% of the full validation sample. Furthermore, as overall in-hospital mortality risk was low in our setting, only 105 in-hospital deaths were observed in this subsample, which decreased the power of the analysis and somewhat limits the interpretation of the calibration plot due to higher uncertainty, especially in the lower decile groups, in which only few events were observed. For both models, we excluded patients with missing information on at least one of the predictors (except mode of arrival), as this information was missing only for a small number of stroke patients. "
A discussion on the study limitations has been provided.

>> Response: We thank the reviewer for approving our discussion of the limitations.

Thanks for providing a copy of the TRIPOD checklist.

>> Response: In line with the editorial suggestion, we have now updated the checklist and provide the newest version of the checklist (TRIPOD+AI) with updated page numbers.

VERSION 2 - REVIEW

Reviewer	**2**
Name	**Kelson, Zoe**
Affiliation	**University of Exeter, Mathematics**
Date	**27-Apr-2025**
COI	

Thanks to the authors for responding to each reviewer comment in turn, providing clarification, undertaking additional analyses, and revising the manuscript where required. The article is much strengthened thanks to the authors' amendments.