PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	GARFIELD-AF Model for Prediction of Stroke and Major Bleeding in Atrial Fibrillation: A Danish Nationwide Validation Study
AUTHORS	Dalgaard, Frederik; Pieper, Karen; Verheugt, Freek; Camm, John; Fox, Keith; Kakkar, Ajay K; Pallisgaard, Jannik; Rasmussen, Peter Vibe; Weert, Henk van; Lindhardt, Tommi Bo; Torp-Pedersen, Christian; Gislason, Gunnar Hilmar; Ruwald, Martin H; Harskamp, Ralf

VERSION 1 – REVIEW

REVIEWER	Jianhua Wu University of Leeds
REVIEW RETURNED	29-Aug-2019

GENERAL COMMENTS	This study is to externally validate the accuracy of the GARFIELD-AF model in a Danish cohort. The paper has clearly defined the risk scores and outcome measures in the validation cohort and the methodology is sound and appropriate. It has to be noted that the prediction accuracy of GARFIELD-AF risk model is quite limited with highest C-index of 0.71, and only improved slightly compared with CHA2DS2VASc score.

REVIEWER	Kui Hong Department of Cardiovascular Medicine, the Second Affilated of Hospital, Nanchang University,China
REVIEW RETURNED	30-Aug-2019

GENERAL COMMENTS	The authors tested the accuracy of the GARFIELD-AF model against existing risk scores for stroke and major bleeding risk in patients with non-valvular atrial fibrillation (AF) in the Danish cohort. They compared the performance with the CHA2DS2-VASc and HAS-BLED for risk of ischemic stroke and major bleeding, respectively. the results showed GARFIELD-AF model might be more effective than CHA2DS2-VASc for the prediction of stroke with a bigger number of sample size. I have some concerns shown as below 1. What the cutoff values for GARFIELD-AF risk means that the benefits outweigh the risks? This point did not show in the paper. 2. It would be more acceptable if the authors provided the net reclassification improvement of stroke risk for GARFIELD-AF model compared with CHA2DS2-VASc. 3. It will be appreciated if the authors provide the data from the comparion between the proportion of patients and event rates for stroke in GARFIELD-AF and CHA2DS2-VASc risk scores (e.g. stratified by low risk, moderate risk, high risk)

	4. For the prediction of low-risk patients with ischemic stroke, both tools have a poor predictive performance (no C-statistic was over 0.70) in the Danish cohort. This point should be emphasized and further discussed.

REVIEWER	Pasquale Pignatelli Sapienza university
REVIEW RETURNED	03-Sep-2019

GENERAL COMMENTS	In this study the authors aimed at externally validate the accuracy of the GARFIELD-AF model against existing risk scores for stroke and major bleeding risk in patients with non-valvular atrial fibrillation (AF) in a population-based cohort. They concluded that In a nationwide Danish cohort with non-valvular AF, the GARFIELD-AF model adequately predicted the risk of ischemic stroke/SE and major bleeding. It is opinion of this referee that the hottest point in the study is represented by the evidence that The GARFIELD-AF model is as efficient as CHA2DS2VASc for stroke and HAS-BLED for major bleeding in atrial fibrillation. Unfortunately This validation is based on ICD-10 coding from the Danish registries which is prone to misclassification bias and lacked clinical measurements. This is a nice study. I wonder if it would be possible to perform an adjunctive analysis on a population from the Garfield study in the sud of Europe in order to confirm the validity of the Garfield score also in subjects from Sud of Europe. Moreover, as af is associated to increase risk for myocardial infarction it would be of interest to evaluate the ability of the Garfield score to predict also in this setting.

REVIEWER	Tuomas Kiviniemi Brigham and Women's Hospital Harvard Medical School Boston, MA, USA
REVIEW RETURNED	16-Sep-2019

GENERAL COMMENTS	Authors sought to externally validate the accuracy of the GARFIELD-AF model against existing risk scores for stroke and major bleeding risk in patients with non-valvular atrial fibrillation (AF) in a large Danish population-based cohort. The topic is relevant and clinically meaningful and text is well written. There is a need for better yet simple-to-use prediction tools for clinical practice. In this sense manuscript assess an important question. Results however, are not a game changer. I have following concerns: 1) Clinical implication of the results remain low. Based on AUC values, GARFIELD score does not add much to the prediction of strokes/major bleeds. 2) Since the main message of paper is better performance of GARFIELD score, more detailed analysis is warranted. Please break down the statistical analysis in more detail when comparing AUCs. Please also provide Net reclassification improvement values.

	3) Limitation section states "...prone to random misclassification bias". It is not really 'random' bias but consists of treating physicians discretion etc as well . Rephrase. 4) Present also score-related data in more detail. Low vs high. 5) Results section should contain more data on comparisons.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
This study is to externally validate the accuracy of the GARFIELD-AF model in a Danish cohort. The paper has clearly defined the risk scores and outcome measures in the validation cohort and the methodology is sound and appropriate. It has to be noted that the prediction accuracy of GARFIELD-AF risk model is quite limited with highest C-index of 0.71, and only improved slightly compared with CHA2DS2VASc score.

- *Response:* We thank the reviewer for these comments. We agree that the improvement in discriminatory ability is slightly improved with GARFIELD-AF over CHA_2DS_2-VASc.

Reviewer: 2
Comment #1. What the cutoff values for GARFIELD-AF risk means that the benefits outweigh the risks? This point did not show in the paper.

- *Response #1.* The GARFIELD-AF model is not categorized but instead provides risk prediction on a continuous scale. The fact that the GARFIELD-AF model is not categorize means that one can make more informed decisions when weighing risks versus benefits. We would argue that it is the role of country and regional guideline committees to determine appropriate cut-points for risk to guide in treatment decisions. In until that guidance is provided, we hope clinicians will use these simultaneous measures of a patients' risk to determine the best treatment. We have changed the discussion to make this more clear

- Discussion (p. 11, line 1 para): *The GARFIELD-AF model is not categorized into risk groups but instead provides risk prediction on a continuous scale The GARFIELD-AF model provides and provides risk estimates of the risk of stroke/SE and bleeding (and mortality when blood pressure and heart rate data are available) in a single calculation and based on routinely collected data would have potentially wide clinical applications.*

Comment #2. It would be more acceptable if the authors provided the net reclassification improvement of stroke risk for GARFIELD-AF model compared with CHA2DS2-VASc.

- *Response #2.* We did not provide net reclassification improvement or stroke risk as this statistic is not appropriate when using it for point-based scores, such as CHA2DS2-VASc. The underlying rationale is explained by Thomas LE together with Pencina MJ (who developed the original NRI measure), et al [1]. We have added this to the method section.

- Discussion (p 13, 1 para): *We did not asses net reclassification improvement as this statistic is not appropriate when using it for point-based scores, such as CHA_2DS_2-VASc* [1].

Comment #3: It will be appreciated if the authors provide the data from the comparion between the proportion of patients and event rates for stroke in GARFIELD-AF and CHA2DS2-VASc risk scores (e.g. stratified by low risk, moderate risk, high risk).

- *Response #3*: As mentioned in our response #1, it was our intention to provide a risk continuum instead of risk categories to allow for more personalized risk/benefit decision making. We would argue that with risk categories (i.e. tertiles of risk) we lose granularity, as well as generating arbitrary cut-off estimates.

Comment #4: For the prediction of low-risk patients with ischemic stroke, both tools have a poor predictive performance (no C-statistic was over 0.70) in the Danish cohort. This point should be emphasized and further discussed.

- *Response #4:* We agree with the reviewer that this point should be emphasized. We have therefore added to the discussion that both prediction models performed rather poorly in those with low thromboembolic risk (CHADSVASC <2). The low rates of events in this sub-population makes accurate prediction difficult.

- *Discussion (p.* 12, para 2): *In our study although both models had modest predictive discrimination for low risk patients, we found that in these patients the GARFIELD-AF model was well calibrated and provided a better prediction than CHA_2DS_2-VASc score in indicating which patients are truly at low-risk for subsequent stroke.*

Reviewer: 3

Comment #1: This is a nice study. I wonder if it would be possible to perform an adjunctive analysis on a population from the Garfield study in the sud of Europe in order to confirm the validity of the Garfield score also in subjects from Sud of Europe. Moreover, as af is associated to increase risk for myocardial infarction it would be of interest to evaluate the ability of the Garfield score to predict also in this setting.

- *Reponse #1:* The reviewer raised the valid point of extrapolation to other populations (than northern Europe). Our validation study is restricted to the Danish population. Secondly, the aim of the paper was to externally validate the GARFIELD-AF risk score for bleeding and stroke. Although of interest, the GARFIELD-AF model does not yet predict risk of MI.

Reviewer: 4

Comment #1 Clinical implication of the results remain low. Based on AUC values, GARFIELD score does not add much to the prediction of strokes/major bleeds.

- *Response #1* We thank the reviewer for this comment. While GARFIELD does not provide perfect prediction, our study found that does have better discriminatory abilities in predicting stroke than the currently recommended standard risk score (CHA_2DS_2-VASC) and allows for simultaneous prediction of bleeding risk. The clinical implications of the difference in AUC of 0.71 (95% CI: 0.70-0.72) for GARFIELD and AUC 0.67 (95% CI: 0.66-0.68) for CHA_2DS_2-VASc can be debated - i.e. what is a 'clinical meaningful' increase in AUC? We would argue that this difference is clinical meaningful and believe that this is emphasized sufficiently in the manuscript.

Comment #2 Since the main message of paper is better performance of GARFIELD score, more detailed analysis is warranted. Please break down the statistical analysis in more detail when comparing AUCs. Please also provide Net reclassification improvement values.

- *Response #2:* We thank the reviewer and would like to emphasize the statistical analysis. As there was no censoring issue because the data did not contain any loss to follow-up within the first year of the study period, the nonparametric approach of the DeLong method of AUC comparison could be used [2]. The DeLong method also avoids any normality assumptions. As for the request for net reclassification improvement we would like to refer to reviewer #2, response #2.
- *Method (p 8, statistical analysis*): *As there was no censoring issue because the data did not contain any loss to follow-up within the first year of the study period, the significance test of C-index differences was tested for significance with DeLongs method* [2].

Comment #3 Limitation section states "...prone to random misclassification bias". It is not really 'random' bias but consists of treating physicians discretion etc as well . Rephrase.

- *Response #3:* We agree with the reviewer that “random” is not the correct word. We have changed this to “non-systematic”.

Comment #4 Present also score-related data in more detail. Low vs high.

- *Response #4:* As the GARFIELD score does not categorize patients in ‘high-risk’ or ‘low-risk’ but rather risk estimates on a continuous scale we do not see how we can present this data that the reviewer requests. Also, according to the TRIPOD checklist for model validation, there are no section on score-related data of low vs. high risk. What we have currently in the manuscript is quite substantial score-related data. For the total Danish cohort, the OAC treated cohort, and the low risk of stroke cohort, we have the number of events, cumulative incidence, AUC curves, and calibration plots as stated in the TRIPOD checklist for model validation. The data for external validation in these three distinguished Danish cohorts are available in the manuscript.

Comment #5 Results section should contain more data on comparisons.

- *Response #5:* Although the manuscript does include comparisons with GARFIELD-AF population on demographics and predictors, we agree that result section should also include comparison on outcomes as stated in the TRIPOD checklist for model validation section 13c. Therefore, we have chosen to expand Table 2 to include number of outcomes for GARFIELD-AF total population and GARFIELD-AF Scandinavian population and presented in the result section.
- *Results (p 9, clinical outcomes): For the GARFIELD-AF Global the rates of events were lower for stroke/SE (1.2%), major bleeding (1.1%), and deaths (4.7%). Similar rates the GARFIELD-AF Scandinavia the proportion of events were lower for stroke/SE (1.0%), major bleeding (1.7%), and deaths (3.7%).*

Table 2. Number of events and deaths for one 1-year follow-up in the Danish population (all patients, patients on OAC, low risk patients), GARFIELD-AF Global, and GARFIELD-AF Scandinavia.

	Danish AF cohort (n=90693)	Patients treated with OAC (n=51180)	Low risk patients (n=20673)	GARFIELD Scandinavia (n=2396)	GARFIELD-AF Global (n=52080)
Ischemic Stroke/SE	2094	994	139	24	599
Major bleeding/Haemorrhagic stroke	2642	1492	242	28*	341⊥
Deaths	10915	4521	623	88	2459

Table 2 footnotes. Low risk patients were defined as CHA_2DS_2-VASc score (≤2 for women, 0-1 for men and >2 for women and >1 for men). Abbreviations: OAC; oral anticoagulation. *Of those treated with OAC (n=1631), ⊥Of those treated with OAC (n=31712).

References

1. Thomas LE, O'Brien EC, Piccini JP, D'Agostino RB, Pencina MJ. Application of net reclassification index to non-nested and point-based risk prediction models: a review. Eur Heart J. 2019 Jun 14;40(23):1880–7.

2. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837–45.

VERSION 2 – REVIEW

REVIEWER	Kui Hong Department of Cardiovascular Medicine, the Second Affiliated Hospital of Nanchang University, Nanchang of Jiangxi, China
REVIEW RETURNED	05-Oct-2019

GENERAL COMMENTS	the revised manuscritp adequately address the reviewer's points. No sepcial comment.

REVIEWER	Pasquale Pignatelli Sapienza University, Italy
REVIEW RETURNED	04-Oct-2019

GENERAL COMMENTS	the authors answered my questions

REVIEWER	Tuomas Kiviniemi Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
REVIEW RETURNED	11-Oct-2019

GENERAL COMMENTS	Authors have adequately addressed my points. There is a missing word in the revised text of Table 2 footnote.