PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Improved risk-stratification of patients with atrial fibrillation: an integrated GARFIELD-AF tool for the prediction of mortality, stroke and bleed in patients with and without anticoagulation
AUTHORS	Fox, Keith; Lucas, Joseph; Pieper, Karen; Bassand, Jean-Pierre; Camm, John; Fitzmaurice, David; Goldhaber, Samuel; Goto, Shinya; Haas, Sylvia; Hacke, Werner; Kayani, Gloria; Oto, Ali; Mantovani, Lorenzo; Misselwitz, Frank; Piccini, Jonathan; Turpie, Alexander G. G.; Verheugt, Freek; Kakkar, Ajay

VERSION 1 – REVIEW

REVIEWER	Rui Oliveira ULS Guarda, Portugal
REVIEW RETURNED	17-Apr-2017

GENERAL COMMENTS	Summarizing the manuscript to a short sentence: The authors propose a new tool to assess which low-risk patients with atrial fibrillation would benefit from oral anti coagulation COMMENTS - Clinical decision about anti coagulation in low-risk patient is demanding and ambiguous. Therefore, this work could be a great help to daily practice. - Discriminatory value, using C-statistics can be classified as good. When applied to low-risk patient discriminatory value decreases. However, still performs better than CHA2DS2-VASc/HAS-BLED ABSTRACT - It would be easier to read if C-statistics values be mentioned always in the same order "mortality, stroke, bleeding" or "stroke, mortality, bleeding" INTRODUCTION - Patients number don´t match reference 4 " ...(for example, in the derivation of CHA2DS2-VASc, 100 patients had a score of 0, and 164 patients a score of 1 out of 1084 patients evaluated by Lip and colleagues 2010).4 5 ", should be corrected " (for example, in the derivation of CHA2DS2-VASc, 103 patients had a score of 0, and 162 patients a score of 1 out of 1084 patients evaluated by Lip and colleagues 2010).4 5 " - "... observed that between 35% (2010-2011) and 51% (2015-2016)..." the reference 6 sampled patients between 2010 and 2015, so didn't reached 2016. The figure 2 from the same reference shows that in the cohort 4 for patients with a CHA2DS2-VASc score =1, around 60% were anti coagulated.

	MATERIAL AND METHODS - "...that occurred within 1 year of enrolment": is 1 year correct? Reference 7 and "Study procedures and outcome measures" describes a 24 months follow-up. Which is correct? - page 10 "The original model for ischaemic stroke/thromboembolism contained sufficiently few factors to be potentially used as a web-based tool without simplification..." what do you mean with "original model"? is the same as full risk model? RESULTS - Header in table 1, page 20 and 21 are missing DISCUSSION - Page 29 "The populations used to derive CHA2DS2-VASc, for example, contained very few low-risk patients (100 patients had a CHA2DS2-VASc score of 0, and 164 patients a score of 1).16", isn't reference 4? - It's possible to anticipate which cutoff values for GARFIELD-AF risk means that the benefits outweigh the risks? or how further studies will assess this theme?

REVIEWER	Rami Riziq Yousef Abumuaileq M. D, Ph. D Former doctorate researcher in Cardiology, Santiago de Compostela University- Spain. Head of Cardiology departments, Palestinian Medical Services- Gaza- Palestine.
REVIEW RETURNED	21-Apr-2017

GENERAL COMMENTS	After making appraisal review of the manuscript with a title: Improved risk-stratification of patients who may not benefit from anticoagulation: an integrated GARFIELD-AF tool for the prediction of mortality, stroke and bleed in atrial fibrillation. ID: bmjopen-2017-017157. We have comments to be taken into accounts by the authors. 1- The readers can understand from the title of the manuscript that the main aim of the work is to develop a leading risk score to predict ischemic stroke, bleeding and mortality at the same time. However, in the first paragraph of the introduction section the authors focus on the ability of the risk score (i.e., CHA2DS2-VASc) to define truly low risk patients. We think that defining patients at truly low stroke risk is another dilemma which needs different approach and not strongly related to the title of the manuscript. The introduction needs to be written in a more coherent manner. The arguments and justifications to make the study should be clear and strong. 2- In the introduction section, Low stroke risk patients (CHA2DS2VASc 0, 1, or 2) are commonly anticoagulated. Patients with CHA2DS2-VASc = 2 could not be considered low risk. The classification and categorization of CHA2DS2-VASc should be reviewed carefully by the authors. This is a very important point. The ESC 2016 guidelines give class IIa indication to anticoagulate male patients with CHA2DS2 VASc=1 and women with CHA2DS2 VASc=2, this means that this group of patients are not at truly low risk of stroke. Across the manuscript the categorization and definition of patients according to CHA2DS2VASc should be uniformed. 3- In table 1, different variables were repeated like carotid disease and cirrhosis. Variable like new onset AF should be defined as it

	constitutes 45% of the overall cohort. 4- The authors continue to criticize the ability of CHA2DS2VASc to discriminate truly low risk patients, while in the result section: The 1 year Kaplan Meier event rates by CHA2DS2VASc score are given in Figure 1. Event rates were rare in the low risk cohort: 1.3% all cause mortality, 0.5% ischaemic stroke/thromboembolism and 0.5% with haemorrhagic stroke/major bleed. Event rates of 0.5% in the low risk category according authors (i.e., CHA2DS2VASc 0-1 male, and 0-2 female) might be considered a good performance of the score. The authors stated that: For patients with no risk factors other than gender as identified by the CHA2DS2VASc score, only 4 out of these 1579 patients experienced a stroke. This might mean that the CHA2DS2VASc score was accurate at defining true low risk patient, but the authors do not comment on that. 5- CHA2DS2VASc is best tested on non anticoagulated patients, this point should be given more attention and focus by the authors in the results and the discussion sections. 6- In table 2, the authors compared GARFIELD-AF with low risk category CA2DS2VASc. Why not the authors give a categorized formula of the GARFIELD-AF. 7- It will be better if the authors give a table in the mean text to demonstrate the variables in simplified GARFIELD-AF and how to calculate. 8- The authors compares GARFIELD-AF with CHA2DS2VASc using just the calibration and discrimination by c- statistic, we think this is not enough and we recommend the authors to use more statistical approaches like the association (i.e., the hazard ratio) , the net reclassification improvement and the integrated discrimination improvement. 9- In the discussion section the authors stated: contemporary data from GARFIELD AF showed that the rate of ischaemic stroke/systemic embolism per 100 person years (1.6%) is less than half the expected rate of 3.9% (95% CI 1.7% 7.6% unadjusted for aspirin) reported by Lip et al in 2010. In the GARFIELD AF 65% of patients were anticoagulated and the percent of 1.6% is within the average of anticoagulated AF patients. The reference given to compare is not an original research article to compare a cohort versus cohort, it was a review analysis. 10- We think that the manuscript has a good impact if the authors focus on the ability of the score to predict over all prognosis of AF patients (i.e., all cause mortality and not as a subtitute for the recommended and specified ischemic and bleeding risk scores) as this will help to define early those patients who need more frequent follow up visits. It is difficult to introduce a score to predict bleeding and ischemic events at the same time, score like this will be difficult to be handle in real world practice and will not help the decision making process regarding to anticoagulate or not. 11- A user friendly tool needs to be easy to remember and easy to calculate

REVIEWER	Javier Mariani Hospital El Cruce, Argentina.
REVIEW RETURNED	18-May-2017

GENERAL COMMENTS	Authors report a new tool (GARFIELD-AF) to predict mortality, stroke and bleeding risks among patients with atrial fibrillation, derived from a large cohort study, and compare the performance with the CHA2DS2-VASc risk score. The paper report both a full

	model and a simplified model (suitable for easy use in electronic systems). Performance of predictive models was analysed using C-statistics for predicted risks from multivariate regression models. Using these methods authors concluded that the new (GARFIELD-AF) tool performed better than CHA2DS2-VASc for both overall population and low risk patients. From the clinical perspective, low risk AF population (and particularly low risk of stroke) is the most important to identify, since benefits could not outweigh risks of anticoagulation in this subset of patients. Major Comments: 1- Although the risk scores are aimed to assist clinical decision making for anticoagulation in a given patient, and the most relevant outcomes to consider are the embolic and bleeding risk, the greater differences between tools were in all-cause mortality. There no comment on this discrepancy. 2- There is no formal comparison between C-statistics (all comparisons are based on the confidence intervals with substantial overlap between both prediction tools), particularly in low risk patients. P-values for the C-statistics comparisons should be presented. 3- It should be emphasized that, for ischaemic stroke/thromboembolism and bleeding prediction, both tools have a poor predictive performance (there is no C-statistic >0.70). Only for mortality the new tool has a C-statistic that could be considered as "good". 4- The difference between C-statistics for ischaemic stroke/thromboembolism and bleeding prediction, is around 0.05, indicating a low improvement in predictive accuracy. 5- It could be useful to add other metrics to compare both prediction models as suggested elsewhere (Circulation. 2009 May 5; 119(17): 2408–2416), particularly to show how the new tool reclassifies low risk patients. 6- There is no section of "limitations" in the discussion. Minor comment: 1- The number of patients exposed to NSAID or Cox-2 (97%) seems excessive, but no definition of the exposition is given in the paper. 2- Calibration plots in the validation cohort should have confidence intervals as in the case of those from derivation cohort.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Rui Oliveira, ULS Guarda, Portugal
The authors propose a new tool to assess which low-risk patients with atrial fibrillation would benefit from oral anticoagulation.

COMMENTS
Clinical decision about anticoagulation in low-risk patient is demanding and ambiguous. Therefore, this work could be a great help to daily practice. Discriminatory value, using C-statistics can be classified as good. When applied to low-risk patient discriminatory value decreases. However, it still performs better than CHA2DS2VASc or HAS-BLED.

Authors' response:

We thank the reviewer for these comments and agree with the need for better tools to assist clinicians in making decisions regarding anticoagulation for atrial fibrillation.

ABSTRACT
It would be easier to read if C-statistics values be mentioned always in the same order "mortality, stroke, bleeding" or "stroke, mortality, bleeding"

Authors' response:
Thank you for the suggestion. The text in the abstract has been amended as advised.

INTRODUCTION
Patients number don´t match reference 4 "(for example, in the derivation of CHA2DS2-VASc, 100 patients had a score of 0, and 164 patients a score of 1 out of 1084 patients evaluated by Lip and colleagues 2010).4 5 ", should be corrected " (for example, in the derivation of CHA2DS2-VASc, 103 patients had a score of 0, and 162 patients a score of 1 out of 1084 patients evaluated by Lip and colleagues 2010).4,5 "

Authors' response:
Thank you for the correction. The text has been amended as advised.

- "... observed that between 35% (2010-2011) and 51% (2015-2016)..." the reference 6 sampled patients between 2010 and 2015, so didn't reached 2016. The figure 2 from the same reference shows that in the cohort 4 for patients with a CHA2DS2-VASc score =1, around 60% were anticoagulated.

Authors' response:
The text has been amended using the published data [reference 6] for cohorts 1 and 4, respectively.

MATERIAL AND METHODS
"...that occurred within 1 year of enrolment": is 1 year correct? Reference 7 and "Study procedures and outcome measures" describes a 24 months follow-up. Which is correct?

Authors' response:
We confirm that the model was developed based on 39,898 patients enrolled between March 2010 and July 2015 who have at least one year of follow-up data. There is not yet complete data on follow-up for all patients over 2 years.

Comment:
page 10 "The original model for ischaemic stroke/thromboembolism contained sufficiently few factors to be potentially used as a web-based tool without simplification..." what do you mean with "original model"? is the same as full risk model?

Authors' response:
We have amended the text [by removing the word "original"] to make it clear that the model for ischaemic stroke/thromboembolism was also used for the web-based tool without simplification.

RESULTS
Header in table 1, page 20 and 21 are missing Authors' response: The table format has been amended as advised.

DISCUSSION

Comment:
Page 29 "The populations used to derive CHA2DS2-VASc, for example, contained very few low-risk patients (100 patients had a CHA2DS2-VASc score of 0, and 164 patients a score of 1).16", isn't reference 4?

Authors' response:
The duplicated reference has been removed.
It's possible to anticipate which cutoff values for GARFIELD-AF risk means that the benefits outweigh the risks? or how further studies will assess this theme? Authors' response: Thank you for this comment. We considered this point at some length prior to submission of the paper and decided that the threshold for defining when the benefits of anticoagulation out-weighs the risks should be a decision that is made in the light additional factors, including patient preferences, but informed by the improved information from this risk score. Thus, the aim if for the GARFIELD score is to allow decision making to be based on more reliable evidence, but not to mandate an anti-coagulation decision based on a particular threshold. However, we anticipate that guideline groups will with to take a view on such thresholds.

Reviewer 2: Rami Riziq Yousef Abumuaileq Head of Cardiology departments, Palestinian Medical Services- Gaza- Palestine

Comment:
The readers can understand from the title of the manuscript that the main aim of the work is to develop a leading risk score to predict ischemic stroke, bleeding and mortality at the same time. However, in the first paragraph of the introduction section the authors focus on the ability of the risk score (i.e., CHA2DS2-VASc) to define truly low risk patients.

We think that defining patients at truly low stroke risk is another dilemma which needs different approach and not strongly related to the title of the manuscript. The introduction needs to be written in a more coherent manner. The arguments and justifications to make the study should be clear and strong.

Authors' response:
Thank you for this comment. We have modified the title of the manuscript and revised the introduction to address these points. The revised title of this manuscript now reads: "Improved risk-stratification of patients with atrial fibrillation: an integrated GARFIELD-AF tool for the prediction of mortality, stroke and bleed in patients with and without anticoagulation"
We have also made it clearer that our goal is to develop an integrated tool for the prediction of mortality, stroke and bleed in atrial fibrillation so that clinicians can weigh-up the potential benefits of anticoagulation against the hazards of bleeding and avoid the treatment of patients who may not benefit from anticoagulation.

Comment:
We have amended the introduction to explain why the CHA2DS2VASc risk score is not sufficiently robust in defining the population at truly low risk of stroke. This clinically important uncertainty is

reflected in the high proportion of patients with a CHA2DS2-VASc score of either '0' or '1' who are currently prescribed anticoagulation. In this paper, the value of our new integrated GARFIELD-AF score is compared with CHA2DS2VASc both overall and for low-risk populations.
In the introduction section, low stroke risk patients (CHA2DS2VASc 0, 1, or 2) are commonly anticoagulated. Patients with CHA2DS2-VASc = 2 could not be considered low risk. The classification and categorization of CHA2DS2-VASc should be reviewed carefully by the authors. This is a very important point. The ESC 2016 guidelines give class IIa indication to anticoagulate male patients with CHA2DS2 VASc=1 and women with CHA2DS2 VASc=2, this means that this group of patients are not at truly low risk of stroke. Across the manuscript the categorization and definition of patients according to CHA2DS2VASc should be uniformed.

Authors' response:
We agree the ESC 2016 guidelines do suggest a possible benefit for oral AC in patients with at least one stroke risk factor (i.e. CHA2 DS2-VASc score of 1 or more for men, and 2 or more for women). However, the ESC 2016 guidelines recommend oral AC only in patients with in patients with at least two risk factors for stroke i.e. CHA2DS2-VASc risk score of 2 or more in men, and 3 or more in women (2016 ESC). Hence, in line with these new guidelines, we consistently evaluate low-risk as a: CHA2DS2-VASc score of 0 or 1 (men) and 1 or 2 (women) i.e. those patients in whom AC is not recommended by the ESC.

Comment:
In table 1, different variables were repeated like carotid disease and cirrhosis. Variable like new onset AF should be defined as it constitutes 45% of the overall cohort.

Authors' response:
Thank you for this observation. We have added the following definition for new onset/unclassified as a footnote to Table 1: the term used when the type of AF cannot be accurately determined in the short interval between diagnosis of AF and enrolment into the study.

Comment:
The authors continue to criticize the ability of CHA2DS2VASc to discriminate truly low risk patients, while in the result section: The 1 year Kaplan Meier event rates by CHA2DS2VASc score are given in Figure 1. Event rates were rare in the low risk cohort: 1.3% all-cause mortality, 0.5% ischaemic stroke/thromboembolism and 0.5% with haemorrhagic stroke/major bleed. Event rates of 0.5% in the low risk category according authors (i.e., CHA2DS2VASc 0-1 male, and 0-2 female) might be considered a good performance of the score. The authors stated that: For patients with no risk factors other than gender as identified by the CHA2DS2VASc score, only 4 out of these 1579 patients experienced a stroke. This might mean that the CHA2DS2VASc score was accurate at defining true low risk patient, but the authors do not comment on that.
Authors' response:
It is true that CHA2DS2VASc differentiates risk. The data do not imply that it is of no value. The key issue is that despite the low stroke rate (only 4 out of these 1579 patients experienced stroke) the registry shows that about 40% of these patients are being anticoagulated. Thus, clinicians are making the decision to anti-coagulate without fully considering potential benefits against hazards of bleeding. Our aim is to provide clinicians with an integrated tool to help physicians make decisions about treatment based on more robust evidence about risks and potential benefits.

Comment:
CHA2DS2VASc is best tested on non-anticoagulated patients, this point should be given more attention and focus by the authors in the results and the discussion sections.

Authors' response:

If non-anti-coagulated patients had the same characteristics and baseline features as anti-coagulated patients then this would be true. However, the patients who do not receive treatment have different characteristics, including co-morbidities, from those who do receive treatment. Therefore, we feel it is important to develop and test our model on all patients, and include oral anticoagulants as an adjustment factor to account for the change in risk after anticoagulation is used. We have modified paragraph 3 of the discussion.

Comment:
In table 2, the authors compared GARFIELD-AF with low-risk category CHA2DS2VASc. Why do the authors not give a categorized formula for the GARFIELD-AF.

Authors' response:
Rather than provide categorized strata of risk, the advent of smart phones, tablets, and the electronic health records allows for the application of a tool with estimates of mortality risk, stroke/systemic embolism risk and bleeding all from the same tool. The task of guideline groups is to decide on the balance of risk versus benefit for overall populations.

Comment:
It will be better if the authors give a table in the mean text to demonstrate the variables in simplified GARFIELD-AF and how to calculate.

Authors' response:
We have provided this in the appendix. Thus it is available for programmers or statisticians to implement in the electronic health records so that risks can be calculated as part of the patient record or for research purposes.
The authors compares GARFIELD-AF with CHA2DS2VASc using just the calibration and discrimination by c- statistic, we think this is not enough and we recommend the authors to use more statistical approaches like the association (i.e., the hazard ratio), the net reclassification improvement and the integrated discrimination improvement. Authors' response: The Hazard Ratio is based, in part, on the spread of the values of the predictor. That is why hazard ratios for factors like age are commonly presented in terms of 5 or 10 year increase. Thus a highly statistically significant HR can be a value very close to 1, and have limited clinical utility. The spread of the values of the GARFIELD-AF score (continuous values between 0 and 1) and the CHA2DS2VASc score takes (0-9) are quite different so comparison of their HR would not be interpretable. One could standardize the Hazard Ratios but the interpretation may still be hazardous when comparing a categorical and a continuous variable.
Laine Thomas (who is an author on this paper) and Michael Pencina (who is the creator of the NRI and IDI) are submitting a paper in which they will show that these statistics should not be used with a categorical variable such as the CHA2DS2VASc score.
In the discussion section the authors stated: contemporary data from GARFIELD-AF showed that the rate of ischaemic stroke/systemic embolism per 100 person years (1.6%) is less than half the expected rate of 3.9% (95% CI 1.7% 7.6% unadjusted for aspirin) reported by Lip et al in 2010.
In the GARFIELD-AF 65% of patients were anticoagulated and the percent of 1.6% is within the average of anticoagulated AF patients.
Authors' response: The discussion has been modified to make these issues more clear and to take account of the impact of anticoagulation.

Comment:
We think that the manuscript has a good impact if the authors focus on the ability of the score to predict over all prognosis of AF patients (i.e., all-cause mortality and not as a substitute for the recommended and specified ischemic and bleeding risk scores) as this will help to define early those patients who need more frequent follow up visits. It is difficult to introduce a score to predict bleeding

and ischemic events at the same time, score like this will be difficult to be handle in real world practice and will not help the decision making process regarding to anticoagulate or not.

Authors' response:
Thank you for this comment. We agree that the new score is a useful prognosticator of all-cause mortality. This reflects the clinical reality that death is the most common adverse event in patients with a diagnosis of AF.

Comment:
We hope this new element of risk score will encourage a more holistic approach to the management of AF patients as well as aiding decisions on anticoagulation. Such decisions need to take account of realistic estimates of bleeding. These points have been clarified in the discussion (para 3).
A user friendly tool needs to be easy to remember and easy to calculate

Authors' response:
We absolutely agree. By making the equations available (see appendix) to work within electronic systems, we hope that this will facilitate risk evaluation without the need to resort to tables or external calculators.

Reviewer 3: Javier Mariani Hospital El Cruce, Argentina.

Authors report a new tool (GARFIELD-AF) to predict mortality, stroke and bleeding risks among patients with atrial fibrillation, derived from a large cohort study, and compare the performance with the CHA2DS2-VASc risk score. The pape report both a full model and a simplified model (suitable for easy use in electronic systems).
Performance of predictive models was analysed using C-statistics for predicted risks from multivariate regression models. Using these methods authors concluded that the new (GARFIELD-AF) tool performed better than CHA2DS2-VASc for both overall population and low risk patients. From the clinical perspective, low risk AF population (and particularly low risk of stroke) is the most important to identify, since benefits could not outweigh risks of anticoagulation in this subset of patients.

Major Comments

Although the risk scores are aimed to assist clinical decision making for anticoagulation in a given patient, and the most relevant outcomes to consider are the embolic and bleeding risk, the greater differences between tools were in all-cause mortality. There no comment on this discrepancy.
Authors' response: We agree that CHA2DS2VASc score was not designed to assess all-cause mortality and therefore it is not surprising that the GARFIELD-AF model performed significantly better than CHA2DS2VASc score for this outcome. As recommended, we have included this important point to paragraph 4 of the discussion.

There is no formal comparison between C-statistics (all comparisons are based on the confidence intervals with substantial overlap between both prediction tools), particularly in low risk patients. P-values for the C-statistics comparisons should be presented.

It should be emphasized that, for ischemic stroke/thromboembolism and bleeding prediction, both tools have a poor predictive performance (there is no C-statistic >0.70). Only for mortality the new tool has a C-statistic that could be considered as “good”. The difference between C-statistics for ischaemic stroke/thromboembolism and bleeding prediction, is around 0.05, indicating a low improvement in predictive accuracy.

Authors’ response:
When reviewing other validations, an increase from 0.64 to 0.67 has been considered a worthwhile improvement in accuracy. However we agree with the reviewer that a c-index of 0.67 is still lower than ideal and in the discussion we refer to the potential added value of other factors including biomarkers. Nevertheless, our aim was to provide a tool based on widely available and routinely collected parameters.

Comment:
It could be useful to add other metrics to compare both prediction models as suggested elsewhere (Circulation. 2009 May 5; 119(17): 2408–2416), particularly to show how the new tool reclassifies low risk patients.

Authors’ response:
We agree that other metrics would be useful. Please see the above comments for reviewer 2.

Comment:
There is no section of “limitations” in the discussion.
Authors’ response: Please see the paragraph at the end of the discussion (page 24).

Minor comment
The number of patients exposed to NSAID or Cox-2 (97%) seems excessive, but no definition of the exposition is given in the paper.

Authors’ response:
We are grateful to the reviewer for pointing out this issue. This variable in the table has been replaced.

Comment:
Calibration plots in the validation cohort should have confidence intervals as in the case of those from derivation cohort.

Authors’ response: The figures have been amended to include the confidence intervals as requested.
Reviewer 4

Comment:
In the manuscript titled "Improved risk-stratification of patients who may not benefit from anticoagulation: an integrated GARFIELD-AF tool for the prediction of mortality, stroke and bleed in atrial fibrillation" Dr. Fox and colleagues use data from 39,898 patients enrolled in the GARFIELD-AF registry to develop risk scores for stroke, mortality and major bleeding. The models were externally validated in the ORBIT-AF registry using data from 10,132 patients with AF. The GARFIELD-AF risk model was specifically evaluated in AF patients considered to be at low risk of stroke. In the internal validation, the GARFIELD-AF risk models for ischemic stroke, mortality and bleeding performed rather well and seem to be well calibrated. Further, in the evaluation of the GARFIELD-AF risk scores in patients considered to be at low risk of stroke (N=7882), the predictive values were attenuated but still better than the CHADSVASc and HAS-BLED risk scores, respectively. In the external validation, the

predictive value was however attenuated, in particular for the bleeding risk model and for ischemic stroke in patients on OAC treatment. Also, in the external validation, the calibration plots indicated overestimation for mortality and stroke, and underestimation of risk for the bleeding risk model.
The paper is well written, the methodology robust, and the data novel. The evaluation of the scores in patients with and without OAC treatment is commendable. The paper would however be further strengthened by addressing the following comments:
It would be valuable to evaluate the performance of the GARFIELD-AF risk scores in the patients considered to be at low risk in the external validation cohort. This would also provide more support to the conclusion.

Authors' response:
We agree that this would have been ideal. Unfortunately, there were too few low-risk patients in ORBIT-AF to allow us to validate in this subgroup. We anticipate that by making this risk score available others will be able to test the performance in large national datasets with the full spectrum of risk.

Comment:
Despite a stringent model derivation, it would be of interest to provide a discussion in the manuscript for several of the variables included in the GARFIELD-AF risk models. For instance, is history of bleeding really an independent predictor for ischemic stroke or only a marker for patients more prone to be without OAC in the GARFIELD-AF registry (and thus at higher risk for stroke)? Surprising that history of bleeding was not among the variables in the GARFIELD-AF bleeding risk model. Vascular disease was among the variables in the GARFIELD-AF bleeding risk model, was this due to confounding with dual or triple antithrombotic therapy in the registry?
Broad global representation is a strength; however, inclusion of geographic region in a risk score may confer specific limitations, is this variable truly generalizable? Is it possible that these factors contributed to the poorer predictive value of the GARFIELD-AF risk model in the external validation concerning ischemic stroke in untreated patients (Supplementary Table 3) and major bleeding risk (Supplementary Table 4)?

Authors' response:
We also were surprised that history of bleeding does not come into the bleeding model. It is possible that vascular disease is a surrogate for the use of antiplatelet therapy (AP), but even with AP in the model, the term "vascular disease" remains significant (p=0.007).

Comment:
The question of inclusion of region versus country versus no location information into the model was one that was debated among the statisticians while working on this project. The consensus amongst the authors was that there are significant differences in the rate of outcomes across regions. We wished to provide some refinement to the estimates so that they are calibrated to the regional averages rather than just using the overall average for everyone.
Adding region did help with discrimination. Region was not a part of the bleeding model and the only regional factor that remained in the ischemic stroke or SE model was the one representing New Zealand, Australia and South Africa. Therefore, the estimate for US was based on an estimate across all of the other countries in the study.
Please discuss the poorer calibration in the external validation. The calibration plots indicate overestimation of stroke and death, and underestimation of major bleeding, in the external validation.

Authors' response:
The reviewer is correct that the calibration of the new scores in the ORBIT population was not as good as in the original cohort. As shown in table 1 of the appendix there were a series of differences in the definition of terms between GARFIELD-AF and ORBIT-AF and the latter included prevalent patients with AF whereas the former included new onset AF, and hence risk characteristics and outcomes differ. In addition, ORBIT AF included too few low risk patients for this important subgroup to be tested. A higher use of anticoagulation was also seen in the ORBIT-AF population.
Consider adding the CHA2DS2VASc model for comparison in Table 4. Authors' response: Since CHA2DS2VASc is designed for stroke, we would prefer to exclude it from a table of mortality.
The comparison with ATRIA provides valuable data, however, for consistency, please add it to the methods section of the manuscript. Have the authors considered comparing the GARFIELD-AF risk score for bleeding with the newer ORBIT bleeding score, at least in the internal validation cohort? Authors' response: We originally intended to validate against the ATRIA and ORBIT scores but the GARFIELD did not collect laboratory values like haemoglobin. Anaemia / haemoglobin were among the strongest factors in the ORBIT score.

Minor comments
Please add the c-indices for the GARFIELD-AF models in external validation data to the abstract.

Authors' response: The C-indices have been added to the abstract as advised (see last sentence of the results section).

Comment:
How was the simplified GARFIELD-AF risk model derived? Authors' response: The eight factors with the largest Wald chi-squares were retained for the simplified model. Model coefficients were then regenerated on this reduced set of factors. For the other two models, few enough factors were retained with the stepwise process so that no further reduction in the number of variables was needed.
3. It will be easier to follow if "systemic embolism" is used consistently throughout the manuscript (instead of "thromboembolism") Authors' response: The text has been amended throughout the manuscript as advised.
4. Likewise, consider using the same terminology for major bleeding throughout the manuscript (e.g. in Introduction; haemorrhagic stroke/major bleeding, in Definition of endpoints; intracranial haemorrhage and any major bleed). Authors' response: The text has been amended to consistently use the term "haemorrhagic stroke" (where appropriate).
5. Please specify the gender cut-offs for low stroke risk patients on page 8, row 31 (similar to other parts in the manuscript) for readers unfamiliar with risk prediction in AF. Authors' response: The gender cut-offs for low stroke risk is defined in the next sentence.
6. How was information concerning outcome events (stroke, bleeding, and mortality) collected (telephone, hospital records, visits)? Authors' response: Outcome events were collected by telephone follow-ups and hospital records.
7. Are any of the results for the different statistical methods presented (coalescent regression, ridge regression, or random forest) (p.13, row 30)? Authors' response: As this publication is aimed at a clinical audience the authors specified that the primary analysis would be performed using of stepwise regression. The outcomes of other statistical the models were very similar (eg coalescent regression). We felt that detailed consideration of the different statistical approaches was outside the scope of this paper and we plan to submit a separate paper, covering all the statistical methods, to a journal with a statistical focus.
8. How were continuous variables handled in the model, how was linearity evaluated? Authors' response: Continuous variables were evaluated using restricted cubic splines. Then the c-indices with the use of these versus including linear terms only was compared to determine whether "simpler" uses of the factors was acceptable.

9. "Any stroke" should probably read "ischemic stroke" (p.28, row 10).)? Authors' response: The text has been amended as advised.
10. Supplementary Table 3 and 4 and Supplementary Figure 1 are, in my opinion, more interesting than Figure 1 and 2 in the manuscript. Please consider switching these. Authors' response: If the editors of BMJ Open feel this is necessary we will do so but prefer the current arrangement in the light of the anticipated readership.

Tables and Figures

Table 1.
There are no p-values in the table as indicated in the footnote. I agree with this and suggest removal of the sentence.

Authors' response: We have now included p values for Table 1.
Figure 2.

Comment:
What was the median predicted risk? Is this stratification clinically meaningful?

Authors' response:
The median 1 year risk for the 3 overall KM figures are 2.7% death, 0.95% ischemic stroke/ SE, 0.92% haemorrhagic stroke or major bleed. The median 1 year risk for the 3 low risk KM figures are 0.92% death, 0.43% ischemic stroke/ SE, 0.35% haemorrhagic stroke or major bleed. This information has been added to the legend. When there is no clear cut-point for risk, median risk is a good alternative for evaluating low/high. It is true that in some cases, overall risk is very low so the differentiation of two distinct risk groups becomes difficult.
Supplementary Table 1.
A) Please clarify "age at time of diagnosis", how does it differ from age at the time of enrolment? Authors' response: In almost every case for GARFIELD, this will be the same thing since patients are enrolled within 6 weeks of diagnosis and age is considered an integer value in this study. If the patient experienced a birthday between the time of diagnosis and enrolment, the difference in age would only be 1 year.
B) Should probably read Medical history of aortic or peripheral artery disease? Authors' response: The table has been amended as advised.
C) What method was used to estimate renal function in the GARFIELD-AF cohort? Authors' response: The presence of renal disease is determined by the clinician and this decision is not supplemented by the collection of laboratory values for kidney function.
Supplementary Table 4. How many patients had complete 3-year follow-up data (row 27)? Authors' response: 2186 patients were followed to 3 years; 1215 died before reaching the 3 year time point. 6348 were censored as alive before the 3 year.

Comment:
References Two references with "accepted" status have been now been published online.

Authors' response:
The references have been updated as advised.

VERSION 2 – REVIEW

REVIEWER	Rui Oliveira UCSP Pinhel, ULS Guarda
REVIEW RETURNED	03-Jul-2017

GENERAL COMMENTS	1 - “To provide an accurate, web-based tool for stratifying patients with atrial fibrillation to facilitate decisions on the potential benefits/risks”. CHA_2DS_2-VASc and HAS-BLED are already user-friendly, they have few parameters, are available on web/app and are easy to interpret according ESC AF Guidelines? 2- “observational studies also demonstrate overuse of OACs in low-risk patients and underuse in high-risk patients in comparison with predicted use based on the CHA2DS2-VASc score and guideline recommendations.25-27”, “The rationale behind this report is to provide clinicians with a more accurate and user-friendly method for stratifying patients according to their risks of death, stroke and major bleeding, and thereby facilitate decisions on prescribing or withholding anticoagulation.”. Although GARFIELD performs, better than CHA_2DS_2-VASc it does not seem to be a great leap forward helping clinicians stratifying which patients benefit from anticoagulation. GARFIELD-study only confirms that patients with a CHA2DS2-VASc score of 0 in men, 1 in women were at very low risks of stroke 3 – Since this report seeks to provide clinicians a tool to stratify patients, in Discussion still lacks some explanation on how this research could be translated to everyday clinicians' practice.

REVIEWER	Rami Riziq-Yousef Abumuaileq Consultant Cardiologist Palestinian Medical Services, Gaza, Palestine Former Ph.D and Postgraduate Fellowship in Cardiac Electrophysiology in University of Santiago de Compostela
REVIEW RETURNED	01-Jul-2017

GENERAL COMMENTS	After making a second revision of the manuscript ID: bmjopen-2017-017157.R1, with a title: Improved risk-stratification of patients with atrial fibrillation: an integrated GARFIELD-AF tool for the prediction of mortality, stroke and bleed in patients with and without anticoagulation. We have to emphasize on: 1- The points which were raised in the first revision and can not be added or modified, then the authors should mention them in the limitation section. 2- The authors should correct and reflect the CHA2DS2-VASc score in a correct way (i.e. truly low risk CHA2DS2-VASc means men with score 0, and women with score 1. Men with score 1 and women with score 2 are at intermediate risk and not truly low risk), and by this they need to to compare the GARFIELD-AF tool with CHA2DS2-VASc in the group of patients at true low risk (i.e men with CHA2DS2-VASc score 0, and women with score 1). 3- More statistical methods are needed to support the result and not just the C. statistics. 4- The GARFIELD-AF score would be more acceptable if the authors emphasized more and to focus on its role at predicting one year mortality and rehospitalizations.

	The authors made a good work which just might need minor modifications to be ready for publications.

REVIEWER	Javier Mariani Hospital El Cruce, Florencio Varela, Buenos Aires, Argentina.
REVIEW RETURNED	25-Jun-2017

GENERAL COMMENTS	Major comment: Most conclusions of the paper are based on comparisons of C-statistics and no p-value for these comparisons are reported. Please, complete these analyses with formal comparisons between C-statistics to make clear that the new score is significantly superior to previous (and commonly used) tools and to support the conclusions of the paper.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Rui Oliveira
Institution and Country: UCSP Pinhel, ULS Guarda, Portugal
Competing Interests: None declared

1 - “To provide an accurate, web-based tool for stratifying patients with atrial fibrillation to facilitate decisions on the potential benefits/risks”. CHA_2DS_2-VASc and HAS-BLED are already user-friendly, they have few parameters, are available on web/app and are easy to interpret according ESC AF Guidelines?

Authors’ response:
In response to the suggestion from reviewer 3 we now provide additional statistical information for the comparison between the GARFIELD-AF risk score and the CHA_2DS_2-VASc score. These results show that the GARFIELD-AF risk score is superior to the CHA_2DS_2-VASc score in predicting mortality and superior to the CHA_2DS_2-VASc score (in predicting stroke/systemic embolism) and the HAS-BLED score (for major bleeding) (results section [page 17 paragraph 3] and Table 3). Further, the integrated scoring system allows the calculation of not only ischaemic stroke/systemic embolism and bleeding risks, but also mortality risk in a single calculator. While we agree that the CHA2DS2-VASc and HAS-BLED scores require relatively few parameters, the integrated GARFIELD-AF score has the potential to work automatically within electronic medical records. For these reasons, we believe that the GARFIELD-AF score has the potential to be of interest to the medical community. These points have been clarified in the manuscript (pages 21 and 22).

2- “observational studies also demonstrate overuse of OACs in low-risk patients and underuse in high-risk patients in comparison with predicted use based on the CHA2DS2-VASc score and guideline recommendations.25-27”, “The rationale behind this report is to provide clinicians with a more accurate and user-friendly method for stratifying patients according to their risks of death, stroke and major bleeding, and thereby facilitate decisions on prescribing or withholding anticoagulation.”. Although GARFIELD performs, better than CHA_2DS_2-VASc it does not seem to be a great leap forward helping clinicians stratifying which patients benefit from anticoagulation. GARFIELD-study only confirms that patients with a CHA2DS2-VASc score of 0 in men, 1 in women were at very low risks of stroke.

Authors’ response:

As noted in response (1) above, the integrated GARFIELD-AF score is superior to the CHA2DS2-VASc and HAS-BLED scores overall and in anticoagulated and non-anticoagulated patients, without the need for two different scoring systems. The integrated GARFIELD-AF score performs better than both the CHA2DS2-VASc and HAS-BLED scores in low-intermediate risk patients and overall in predicting mortality, and at least as well for stroke/systemic embolism and major bleeding (please see revised manuscript page 17, Table 3).

3 – Since this report seeks to provide clinicians a tool to stratify patients, in Discussion still lacks some explanation on how this research could be translated to everyday clinicians' practice.

Authors' response:
We thank the reviewer for this suggestion and have revised the discussion accordingly (please see revision on page 24).

Reviewer: 2
Reviewer Name: Rami Riziq-Yousef Abumuaileq
Institution and Country: Consultant Cardiologist, Palestinian Medical Services, Gaza, Palestine; Former Ph.D and Postgraduate Fellowship in Cardiac Electrophysiology in University of Santiago de Compostela
Competing Interests: None declared

After making a second revision of the manuscript ID: bmjopen-2017-017157.R1, with a title: Improved risk-stratification of patients with atrial fibrillation: an integrated GARFIELD-AF tool for the prediction of mortality, stroke and bleed in patients with and without anticoagulation. We have to emphasize on:

1- The points which were raised in the first revision and cannot be added or modified, then the authors should mention them in the limitation section.

Authors' response:
We thank the reviewer for this suggestion and have expanded the discussion on the limitations accordingly (please see revision on pages 23 and 24).

2- The authors should correct and reflect the CHA2DS2-VASc score in a correct way (i.e. truly low risk CHA2DS2-VASc means men with score 0, and women with score 1. Men with score 1 and women with score 2 are at intermediate risk and not truly low risk), and by this they need to to compare the GARFIELD-AF tool with CHA2DS2-VASc in the group of patients at true low risk (i.e men with CHA2DS2-VASc score 0, and women with score 1).

Authors' response:
We thank the reviewer for this suggestion and we have modified the wording as advised. For patients with no risk factors other than gender as identified by the CHA2DS2-VASc score (classified in this revision as very low risk), only 4 out of these 1579 patients experienced a stroke or systemic embolism and only 3 out of the 685 anticoagulated patients experienced a major bleed, and so further assessment of these patients was not possible. For ease of interpretation, in the revised manuscript (Table 3) we provide the results with two categories of lower risk: "very low to low risk" = CHA2DS2-VASc score of 0 or 1 [men] and 1 or 2 [women]; HAS-BLED 0 for bleeding. "Low to intermediate risk" = CHA2DS2-VASc score 0, 1 or 2 [men] and 1, 2 or 3 [women]; HAS-BLED score 0 or 1 for bleeding.

3- More statistical methods are needed to support the result and not just the C. statistics.

Authors' response:

We thank the reviewer for this suggestion and additional statistical methods and results are now provided. There are two major areas explored in predictive modelling. One is discrimination for which the c-index is the most frequently quoted statistic. The other is calibration, which is often explored through calibration curves. We wish to point to the fact that we had included calibration curves for the models as well. There are several additional tests for nested models, such as the case when one is exploring the addition of a biomarker to an established model. This is less so for non-nested models. A method that has been suggested is to develop the “super” model, which is the combination of the two models being compared and tests if both individual models add in the presence of the other.

We have performed this test and have provided the p-values for these results in Table 3. We hope the reviewer finds this adequately explores the question.

4- The GARFIELD-AF score would be more acceptable if the authors emphasized more and to focus on its role at predicting one year mortality and rehospitalizations.

Authors’ response:
We thank the reviewer for this suggestion. Accordingly, we have modified the manuscript to reflect the fact that the GARFIELD-AF score accurately predicts mortality (see revised manuscript Table 4). Data on hospitalizations were gathered, but are not part of the GARFIELD-AF score.

Comment:
The authors made a good work which just might need minor modifications to be ready for publications.

Reviewer: 3
Reviewer Name: Javier Mariani
Institution and Country: Hospital El Cruce, Florencio Varela, Buenos Aires, Argentina.
Competing Interests: None declared.

Major comment:
Most conclusions of the paper are based on comparisons of C-statistics and no p-value for these comparisons are reported. Please, complete these analyses with formal comparisons between C-statistics to make clear that the new score is significantly superior to previous (and commonly used) tools and to support the conclusions of the paper.

Authors’ response:
We thank the reviewer for this suggestion and additional statistical methods and results are now provided. There are issues with comparing c-indices for non-nested models. A method that has been suggested is to develop the “super” model, which is the combination of the two models being compared and tests if both individual models add in the presence of the other. We have performed this test and have provided the p-values for these results in Table 3. We hope the reviewer finds this adequately explores the question.

VERSION 3 – REVIEW

REVIEWER	Rui Oliveira ULS Guarda, POrtugal
REVIEW RETURNED	03-Aug-2017

GENERAL COMMENTS	Good improvements were made. Results: - "or low to intermediate risk for CHA2DS2-VASc (0, 1, 2 for men and 1, 2, 3 for women). Higher risk was classified as CHA2DS2-VASc score ≥2 for men and ≥3 for women". Men at 2 and women at 3 are high risk

REVIEWER	Rami Riziq-Yousef Abumuaileq Palestinian Medical Services Gaza Palestine M.D, PhD Cardiology, University of Santiago de Compostela Spain
REVIEW RETURNED	04-Aug-2017

GENERAL COMMENTS	After making the third review of the Manuscript ID bmjopen-2017-017157.R2, entitled "Improved risk-stratification of patients with atrial fibrillation: an integrated GARFIELD-AF tool for the prediction of mortality, stroke and bleed in patients with and without anticoagulation." We still have some comments, 1- In table 3, men with CHA2DS2VASc = 2 and women with CHA2DS2VASc = 3 are at high risk and can not be considered intermediate risk. The authors should define and use CHA2DS2VASc score as it was perceived (i.e in the correct way) and in the methodology section they can explain how they will use the CHA2DS2VASc score in such way. 2- It will be appreciated if the authors include a table in the main text of the manuscript showing the association between individual components of simple GARFIELD-AF tool and the outcomes (i.e stroke, bleeding and mortality). 3- If a patient is at high risk according GARFIELD-AF tool, this means that he is at high risk of having stroke and bleeding at the same time and this could be confusing. So the authors need to justify the usefullness of their tool regarding this high risk category of patients (i.e the can explain that as by this tool, they can define those high risk patients who need close and more follow up visits to minimize their risk of both stroke and bleeding). 3- The limitation section should include at least all of the possible major limitations of the study as this will give strength and more respect to the study. Moreover, the authors should state in the last paragraph before the conclusions that future large cohort studies are needed to confirm the results obtained by the authors.

VERSION 3 – AUTHOR RESPONSE

Reviewer 1

Results: "or low to intermediate risk for CHA2DS2-VASc (0, 1, 2 for men and 1, 2, 3 for women). Higher risk was classified as CHA2DS2-VASc score ≥2 for men and ≥3 for women". Men at 2 and women at 3 are high risk.

Authors' response:
Please can we respectfully refer to the stroke rates associated with the CHA2DS2-VASc score (see table and reference below) to point out that there is not a simple "cut point" for high risk, rather this is a continuous scale. This is the rationale for our use of the terms "very low, low risk, intermediate risk and higher risk"

Annual stroke risk

CHA2DS2-VASc score	Stroke risk %
0	0
1	1.3
2	2.2
3	3.2
4	4.0
5	6.7
6	9.8
7	9.6
8	12.5
9	15.2

Hohnloser SH, Duray GZ, Baber U, Halperin JL. Prevention of stroke in patients with atrial fibrillation: current strategies and future directions Eur Heart J. 2008; 10 (Suppl H): H4–H10. https://doi.org/10.1093/eurheartj/sun029

Reviewer 2
In Table 3, men with CHA2DS2-VASc = 2 and women with CHA2DS2-VASc = 3 are at high risk and cannot be considered intermediate risk. The authors should define and use CHA2DS2-VASc score as it was perceived (i.e. in the correct way) and in the methodology section, they can explain how they will use the CHA2DS2-VASc score in such way.

Authors' response:
Please can we refer to our response to reviewer 1 (above). We believe that by suggesting an arbitrary cut point for "high risk" (CHA2DS2-VASc = 3) this would not reflect the reality of the continuous and progressively increasing risk for stroke over the CHA2DS2-VASc scale and it may lead to some misunderstandings by the readership.

Comment:
It will be appreciated if the authors include a table in the main text of the manuscript showing the association between individual components of simple GARFIELD-AF tool and the outcomes (i.e. stroke, bleeding and mortality).

Authors' response:
We thank the reviewer for this suggestion. We have added supplementary Table 2, which details the Wald Chi-square, p-values and hazard ratios for each component of the simplified GARFIELD-AF models. This table is referenced on page 13 of the manuscript and in the discussion (on page 25).

Comment:
If a patient is at high risk according GARFIELD-AF tool, this means that he is at high risk of having stroke and bleeding at the same time and this could be confusing. So the authors need to justify the usefulness of their tool regarding this high risk category of patients (i.e. they can explain that as by this tool, they can define those high risk patients who need close and more follow up visits to minimize their risk of both stroke and bleeding).

Authors' response: We thank the reviewer for this comment and we have accordingly added two sentences to the discussion to indicate that by identifying patients as at high risk of both stroke and bleeding it would be appropriate for such patients to be closely monitored and steps to be taken to modify any potentially reversible bleeding risks. As the GARFIELD risk score is more accurate than HAS-BLED in identifying those at higher bleeding risk, this provides a more reliable basis for subsequent clinical decisions. We have added a couple additional sentences on page 23 of the Discussion.

Comment:
The limitation section should include at least all of the possible major limitations of the study as this will give strength and more respect to the study. Moreover, the authors should state in the last paragraph before the conclusions that future large cohort studies are needed to confirm the results obtained by the authors.

Authors' response: We thank the reviewer for this comment and we have further expanded the limitations section of the manuscript (please see revisions on pages 25 and 26). In addition, as suggested by the reviewer, prior to the conclusion (on page 26), we have suggested that large cohort studies should be used to confirm the findings of this study.