Improving medical decision making: Stroke prevention in atrial fibrillation
Arts, D.L.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 8

Development of a regression model and neural network to predict stroke and comparison to the CHA$\text{2}DS_2$-VASc in a large dataset

D.L. Arts, H.A. van den Ham, T. van Staa, H.C.P.M. van Weert, A. Abu-Hanna
ABSTRACT

Aims
International guidelines for atrial fibrillation use the CHA2DS2-VASc to determine anticoagulation strategies. This risk score performs poorly at classifying low risk patients, leading to overtreatment which should be avoided due to the risk of adverse events and increased healthcare cost. We aim to improve stroke risk prediction by using logistic regression and neural networks. We will investigate if patients can benefit from more precise predictions and thus more tailored treatment for atrial fibrillation. This in turn should result in reduced overtreatment of low risk patients.

Methods and Results
We trained a logistic regression (LR) model and a neural network (ANN) to predict stroke in a large cohort without anticoagulant treatment. We compared stroke predictions by these models with CHA2DS2-VASc based predictions. Low and high risk groups were created to compare performance in these groups. The Net Reclassification Index (NRI) was calculated to investigate the number of overtreatment cases that could be avoided by using prediction models. The ANN performed best (c-index: 0.67 95% CI [0.63,0.70]), followed by the LR model (c-index: 0.66 95% CI [0.63,0.70]) and the CHA2DS2-VASc (c-index: 0.64 95% CI [0.60,0.67]). The ANN reduced overtreatment of low risk patients with 59%, followed by the LR with 56%. Using the ANN for stroke prediction could have saved 12 intracerebral hemorrhage and 58 major bleeding events per 100,000 patients respectively. Undertreatment increased less than 1% for both models.

Conclusion
Logistic regression models and neural networks can improve stroke risk prediction and reduce overtreatment when compared to the CHA2DS2-VASc.
Development of a regression model and neural network to predict stroke

INTRODUCTION

Patients with atrial fibrillation (AF) are at increased risk for stroke.[1] Oral anticoagulants (OAC) such as warfarin and the new oral anticoagulants (NOACs), can reduce this risk, but may have serious side-effects[2, 3]. Determining which patient should receive this medication has been a topic of discussion for many years. Most recent guidelines recommend prescribing OAC for all but the lowest risk categories to allow for easy implementation and optimal stroke risk reduction. To determine which patients should receive OAC, stroke risk scores are used to predict individual risk of stroke for each patient and balanced against possible side effects. Several such schemes have been developed in the past and these are used in international guidelines.[4, 5]

Currently, the CHA2DS2-VASc is the most widely used stroke risk score: it consists of 7 variables that can add up to a total of 9 points and generally performs better at predicting stroke than the CHADS2.[4] Due to the high net benefit of treatment with antithrombotic medication, the stroke risk cut off for this treatment has been lowered over the past few years.[6] The most recent ESC guidelines recommend treatment for patients scoring 1 point or more. The result is that almost any patient over 65 years old with AF should receive antithrombotic medication. While this treatment strategy might be optimal for most AF patients, it results in overtreatment for some, namely low risk patients. The issue here is the coarseness of stroke risk scores.[7, 8] This lack of precision inevitably results in guidelines recommending treatment for low risk patients that is not properly tailored to their risk profile. A secondary issue with stroke prevention in AF is low guideline adherence. Several studies have found that a large proportion of patients is not treated in accordance with recent guidelines.[9, 10] While we recognize that this topic is of the utmost importance, the scope for this study is limited to improving stroke prediction accuracy.

Statistical methods like logistic regression (LR) and artificial neural networks (ANN) have been around for many years and play an important role in certain fields of medicine.[11] For example, ANN are currently used for the prediction of breast cancer and analysis of MRI images.[12, 13] Furthermore these methods have have already been put to use in the prediction of AF from ECGs, showing excellent accuracy.[14] Overall, however, its use is rare due to a lack of expertise and experience. We hypothesize that the increasing number of stroke risk factors allows for more accurate predictions when used in models than in simple additions like the CHA2DS2-VASc, as these models can better mimic the complex real life associations between risk factors.

We aim to investigate if stroke risk prediction can be improved by developing a logistic regression model and neural network. We will investigate if patients can benefit from more precise predictions and thus more tailored treatment for atrial fibrillation. This in turn should result in reduced overtreatment of low risk patients.
METHODS

Figure 1 contains a visual overview of analyses performed for this study: We trained a logistic regression model and a Neural Network and compared their predictions for stroke with CHA_{2}DS_{2}-VASc based predictions. Low and high risk groups were created to compare performance in these groups.

**Figure 1.** Structured overview of analyses performed.

**Data source**

The Clinical Practice Research Datalink contains anonymized information of over 11 million patients registered at over 600 general practices in the United Kingdom from 1998-2012. Information is continuously recorded for each patient, including a record of each consultation, diagnoses, prescribed medicines, and basic demographic data. The geographical distribution and size of general practices represented in the Clinical Practice Research Datalink is largely representative for England and Wales, and people
Development of a regression model and neural network to predict stroke

registered in the database are representative of the UK population in terms of age and sex. The quality of data is subject to rigorous checks and regular audits.[15, 16] About 50% of the practices in CPRD have been linked to other datasets in England such as the Hospital Episode Statistics (HES). The HES includes records of inpatient hospitalizations, such as date of admission and discharge, diagnoses and procedures done. The current study only used the practices that have been linked to HES to ensure high quality data. Risk factor were based on diagnostic codes unless otherwise specified below. Diagnostic codes were extracted from both HES (ICD-10 codes) and CPRD (Read-codes). For renal failure and proteinuria, lab test records were used in addition to diagnostic codes. Renal dysfunction was defined as an estimated glomerular filtration rate (eGFR) of <60 ml/min/1.73m² or a diagnosis of renal failure. Patients that did not have a lab value or a diagnostic code for renal failure or proteinuria were assumed to have normal values. Smoking was defined as ‘current smoker’. All risk factors were assessed at baseline.

**Study population**
Patients aged ≥18 years with a record of AF and no prior warfarin/NOAC or heparin prescription were included. Patients with a record of rheumatic valve disease or valvular repair/replacement were excluded. The first occurrence of AF at least 12 months after the inclusion in the database was used as an index date. The outcome of interest was ischemic stroke recorded either in CPRD (according to Read coding) or HES (according to ICD-10 codes). Patients were censored at either start of warfarin or NOAC use, stroke, death or end of follow up.

**Variable selection**
Variables selection for our prediction model was based on recent literature on predictors for stroke in AF and availability CPRD.
Associations with outcome measure were calculated in a univariate logistic regression model using stroke during follow up as outcome.

**Missing data**
Missing data (BMI, smoking, ethnicity) were imputed with the MICE library for R using Gibbs sampling, to allow for inclusion of all cases into the analyses.[17]

**Risk prediction models**
Logistic Regression (LR) and Artificial Neural Networks (ANN) were used to predict the risk of stroke. Stroke during follow up was used as dependent variable. In the logistic regression model, variable selection was performed using backwards selection based on the Akaike Information Criterion (AIC). The model was trained on a random 85% sample of the dataset (training set) and the remaining 15% was used to
validate the model (validation set). This validation set was stored in a separate location and was only used for final validation.

The main reason for using Neural Networks is the fact that they can automatically model complex non-linear relationships. This can improve predictions when compared to simpler models that pose stringent assumptions (e.g. linearity between covariate and outcome).

We used the R library “NeuralNet” to create and train a neural network. The same training and validation set that were used to train and validate the LR were used for the ANN. The network was initially created with a default setup of 1 ‘hidden layer’ and 1 ‘node’. This initial setup was optimized by way of a ‘grid search’. This approach systematically evaluates the performance of a network architecture by modifying parameters that determine how the network functions.

These models were compared to the CHA2DS2-VASc. The CHA2DS2-VASc uses 7 variables to predict stroke (congestive heart failure, hypertension, age between 65 and 74 years and age >75 years, prior stroke/TIA, vascular disease, diabetes mellitus, and female sex). Patients can score from 0 to 9 points. The European Society of Cardiology recommends OAC (warfarin/NOACs) for patients scoring 1 point or more.[4, 18, 19]

Both methods used the same initial variables. The LR model and ANN were built using R.[20]

**Low and high risk groups**

As net benefit of antithrombotic medication for patients at high risk for stroke is great, most room for improvement is expected to be in low risk patients. To separate patients into a low risk and a high risk group, we calculated a one year stroke risk using Cox regression (figure 1). AF patients with an estimated 1 year stroke risk below 1.7% were classified as low risk, the others as high risk. The 1.7% stroke rate is based on a study that determined the cut-off annual stroke rate where OAC treatment leads to net benefit.[21]

**Predictive performance analyses**

To compare predictive performance, the CHA2DS2-VASc was used as a continuous variable in a logistic regression model with stroke during follow up as outcome. Likewise the continuous predicted values from the LR model and the ANN were used. The C-statistic, weighted for follow up time, was used to compare the discriminating ability of each method.[22] Furthermore, the net reclassification index (NRI) was calculated to provide a more clinical and easier to interpret measure of model improvement. The Brier score was calculated to assess the accuracy of the predictions. It measures the mean squared difference between the predicted probability assigned to the possible outcomes and the actual outcome. The lower the Brier score, the better the model accuracy.
Net Reclassification Index

The Net Reclassification Index (NRI) is a measure for evaluating the improvement in prediction performance of prediction models. It reflects the reclassification of cases (event NRI) and non-cases (non-event NRI) as result of a new model when compared to an old model.[23] Although its use has been under discussion[24, 25], it remains one of the easier to interpret measures for prediction performance. The NRI consists of two sub scores: the event NRI and the non-event NRI. The event NRI resembles the number of patients with the outcome of interest (cases) that have been correctly reclassified from non-case to case, when comparing predictions from the old model to the new model. The non-event NRI works the other way; it resembles correct reclassification of cases to non-cases by the new model. The total NRI is a weighted combination of the sub scores. The higher the NRI, the better the new model performs compared to the old model.

The R package “nricens” was used for generating NRIs. To calculate the NRI for the CHA2DS2-VASc, a binary variable of the CHA2DS2-VASc was created, making all CHA2DS2-VASc scores > 0 correspond with a binary value of 1, thus reflecting ESC-guideline recommendations for prescribing OAC for scoring > 0 points. The CHA2DS2-VASc was used as ‘old’ model, and the ANN and the LR model were used as ‘new’ models. Using this approach we ensured we assessed effects pertaining to the actual clinical situation, i.e. whether or not a physician would prescribe OAC for a specific patient.

Reduction in adverse events

To calculate how many adverse events we can prevent by reducing overtreatment, we assume a number needed to harm (NNH) of 909 and 208 for intracranial hemorrhage (ICH) and major bleeding events respectively (based on the 2007 study by Hart et. al.)[26]. We can calculate how many events can be prevented by taking the proportion of patients correctly reclassified as non-event, multiplied by the number of patients in the low risk group, divided by the NNH.

RESULTS

A total of 60594 patients was included in our analyses with a total of 125297 patient years and a mean follow up duration of 1 year. Patients were on average 74 years old (SD 12.11), over half was male (51%), approximately 15% had suffered a previous stroke and 6% suffered a stroke during follow up (Table 1). The annual stroke rate was 3%. The mean CHA2DS2-VASc was 3.30 (Median: 3.00; IQR [2.00, 4.00]) with 93% of patients scoring 1 point or more, meaning they should receive OAC according to ESC-guidelines. 1604 patients were classified as low risk (18%) and 7468 patients as high risk (82%) in the validation set.
Table 1. Characteristics of study population demographics

<table>
<thead>
<tr>
<th>Follow up until censoring (years)³</th>
<th>No stroke during follow up</th>
<th>Stroke during follow up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>74.0 (12.2)</td>
<td>79.8 (9.0)</td>
</tr>
<tr>
<td>&lt; 65</td>
<td>11506 (20%)</td>
<td>236 (6%)</td>
</tr>
<tr>
<td>65-75</td>
<td>14245 (25%)</td>
<td>678 (18%)</td>
</tr>
<tr>
<td>75-85</td>
<td>20019 (35%)</td>
<td>1618 (43%)</td>
</tr>
<tr>
<td>85+</td>
<td>11073 (19%)</td>
<td>1219 (32%)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>29640 (52%)</td>
<td>1448 (39%)</td>
</tr>
<tr>
<td>Female</td>
<td>27203 (48%)</td>
<td>2303 (61%)</td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smoker</td>
<td>8913 (16%)</td>
<td>508 (14%)</td>
</tr>
<tr>
<td>Non-smoker or unknown</td>
<td>47930 (84%)</td>
<td>3243 (86%)</td>
</tr>
<tr>
<td>BMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>8134 (14%)</td>
<td>755 (20%)</td>
</tr>
<tr>
<td>&lt; 20</td>
<td>2534 (4%)</td>
<td>202 (5%)</td>
</tr>
<tr>
<td>20-25</td>
<td>14688 (26%)</td>
<td>1037 (28%)</td>
</tr>
<tr>
<td>25-30</td>
<td>18574 (33%)</td>
<td>1177 (31%)</td>
</tr>
<tr>
<td>&gt; 30</td>
<td>12913 (23%)</td>
<td>580 (15%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>44788 (79%)</td>
<td>3059 (82%)</td>
</tr>
<tr>
<td>Other⁴</td>
<td>773 (1%)</td>
<td>53 (1%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>11282 (20%)</td>
<td>639 (17%)</td>
</tr>
<tr>
<td>Social economic status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0% 20% (most deprived)</td>
<td>12877 (23%)</td>
<td>825 (22%)</td>
</tr>
<tr>
<td>21-40%</td>
<td>14656 (26%)</td>
<td>941 (25%)</td>
</tr>
<tr>
<td>41-60%</td>
<td>12145 (21%)</td>
<td>782 (21%)</td>
</tr>
<tr>
<td>61-80%</td>
<td>10194 (18%)</td>
<td>709 (19%)</td>
</tr>
<tr>
<td>81-100% (least deprived)</td>
<td>6971 (12%)</td>
<td>494 (15%)</td>
</tr>
</tbody>
</table>
Variable selection

Variable selection is displayed in Table 2. Gender, age, history of stroke and diabetes were selected for every model. Hypertension was not selected for the LR model and ANN, vascular disease was not selected by the LR and Cox models. Congestive heart failure was only selected for the LR model. Only the LR model selected ethnicity as a relevant predictor. Smoking was selected as a strong predictor for the ANN. Table 3 displays odds ratios for the selected logistic regression model. Figure 2 shows a graphical representation of the ANN.
Prediction and misclassification

In the validation dataset, the ANN performed best (weighted c-index: 0.67 95% CI [0.63,0.70]), followed closely by the LR model (weighted c-index: 0.66 95% CI [0.63,0.70]) and the CHA2DS2-VASc (weighted c-index: 0.64 95% CI [0.60,0.67]) (table 4). These differences were not statistically significant. All models had a Brier score of 0.06.

Table 5 shows the Net Reclassification Index. On the overall NRI, the LR model and the ANN performed similarly (NRI: 0.19 95% CI [0.18, 0.2] for the LR and [0.19, 0.2] for the ANN) compared to the CHA2DS2-VASc risk score. In the low risk group, the ANN performed best (NRI: 0.58 95% [0.55, 0.6]) followed by the LR (NRI: 0.56 95% CI [0.54, 0.59]) compared to the CHA2DS2-VASc risk score. The non-event NRI (NRIn) in the low risk group is the most interesting statistic to study as it reflects patients that were accurately down-coded in the low risk group, i.e. reduced overtreatment. The ANN showed an NRIn of 0.59 (95% CI [0.56, 0.61]), the LR model showed an NRIn of 0.57 (95% CI [0.55, 0.60]).

Using the ANN for stroke prediction could have saved 1 ICH events and 5 major bleeding events in the validation group, and roughly 7 ICH events and 35 major bleeding events in the total study population. These prevented events come at no additional cost (economically or harm). This is calculated by taking the proportion of patients correctly reclassified as non-event (0.59) and multiplying that with the number of patients in the low risk group (1604) and finally dividing by the NNH.

Table 2. Variable selection

<table>
<thead>
<tr>
<th>Variables</th>
<th>LR</th>
<th>ANN</th>
<th>COXPH</th>
<th>CHADS2</th>
<th>CHA2DS2-VASc</th>
<th>Univariate LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>Age</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>History of stroke</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>Vascular comorbidity</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>P &lt; 0.01</td>
</tr>
<tr>
<td>Hypertension</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>Diabetes</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>P = 0.25</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.01</td>
</tr>
<tr>
<td>Renal failure</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>Proteinuria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P = 0.24</td>
</tr>
<tr>
<td>Social economic status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P = 0.05</td>
</tr>
<tr>
<td>BMI (categorical)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>Smoking status</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P = 0.30</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P = 0.10</td>
</tr>
</tbody>
</table>

Variables for logistic regression (LR) and Cox proportional hazards models were selected using AIC based stepwise backwards selection. Artificial Neural Network (ANN) input variables were selected using a grid search, i.e. testing each combination of variables on a training set and evaluating performance.

Prediction and misclassification

In the validation dataset, the ANN performed best (weighted c-index: 0.67 95% CI [0.63,0.70]), followed closely by the LR model (weighted c-index: 0.66 95% CI [0.63,0.70]) and the CHA2DS2-VASc (weighted c-index: 0.64 95% CI [0.60,0.67]) (table 4). These differences were not statistically significant. All models had a Brier score of 0.06.

Table 5 shows the Net Reclassification Index. On the overall NRI, the LR model and the ANN performed similarly (NRI: 0.19 95% CI [0.18, 0.2] for the LR and [0.19, 0.2] for the ANN) compared to the CHA2DS2-VASc risk score. In the low risk group, the ANN performed best (NRI: 0.58 95% [0.55, 0.6]) followed by the LR (NRI: 0.56 95% CI [0.54, 0.59]) compared to the CHA2DS2-VASc risk score. The non-event NRI (NRIn) in the low risk group is the most interesting statistic to study as it reflects patients that were accurately down-coded in the low risk group, i.e. reduced overtreatment. The ANN showed an NRIn of 0.59 (95% CI [0.56, 0.61]), the LR model showed an NRIn of 0.57 (95% CI [0.55, 0.60]).

Using the ANN for stroke prediction could have saved 1 ICH events and 5 major bleeding events in the validation group, and roughly 7 ICH events and 35 major bleeding events in the total study population. These prevented events come at no additional cost (economically or harm). This is calculated by taking the proportion of patients correctly reclassified as non-event (0.59) and multiplying that with the number of patients in the low risk group (1604) and finally dividing by the NNH.
DISCUSSION

In this study we aimed to improve stroke prediction in atrial fibrillation overall and for low risk groups. Neural networks performed best on a separate validation set, closely followed by logistic regression compared to the CHA2DS2-VASc risk score, differences in the weighted C statistic were not significant. When examining reclassification, the ANN performed better in the low risk group, although differences with the LR model were marginal.

As the CHA2DS2-VASc classifies most patients in the (moderate to) high risk groups (93%), there is little to be gained in the high risk groups in terms of reclassification. Consequently, much can be gained in the low risk group by downcoding non-events. In the low risk group, the ANN correctly reclassified 59% of the CHA2DS2-VASc cases as a non-case and the LR model 57% of CHA2DS2-VASc cases.

This study shows CHA2DS2-VASc performs poorly for low risk patients, resulting in unnecessary (over)treatment in clinical practice. Relatively poor performance of the CHA2DS2-VASc, and other stroke risk prediction schemes, has been found in several large studies, showing similar c-indices.[7, 27, 28] Other studies report higher c-indices for CHA2DS2-VASc. Differences in the predictive ability can be explained by different type of
Figure 2. The final neural network, colors indicate the relative importance of each variable, red being the most important, green the least. The thickness of each line indicates the (weight of) the influence the nodes have on each other.

Table 5. Net reclassification of CHA₂DS₂-VASc to new models

<table>
<thead>
<tr>
<th>All patients</th>
<th>Base model</th>
<th>New model</th>
<th>NRIe 95% CI</th>
<th>NRIn 95% CI</th>
<th>NRI 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHA₂DS₂-VASc</td>
<td>LR</td>
<td>0</td>
<td>[-0.01, 0]</td>
<td>0.19</td>
<td>[0.19, 0.2]</td>
</tr>
<tr>
<td>CHA₂DS₂-VASc</td>
<td>ANN</td>
<td>-0.01</td>
<td>[-0.01, 0]</td>
<td>0.2</td>
<td>[0.19, 0.21]</td>
</tr>
<tr>
<td>LR</td>
<td>ANN</td>
<td>0</td>
<td>[0, 0]</td>
<td>0.01</td>
<td>[0, 0.01]</td>
</tr>
</tbody>
</table>

| Low risk group (yearly stroke risk < 1.7%) |
|------------------------------------------|-----------|-------------|-------------|-------------|-------------|
| Base model                               | New model | NRIe 95% CI | NRIn 95% CI | NRI 95% CI |
| CHA₂DS₂-VASc | LR         | -0.01     | [-0.01, 0]  | 0.57        | [0.55, 0.6] |
| CHA₂DS₂-VASc | ANN        | -0.01     | [-0.01, -0.01] | 0.59    | [0.56, 0.61]| 0.58       | [0.55, 0.6] |
| LR            | ANN        | 0         | [0, 0]      | 0.01        | [0, 0.02]   | 0.01       | [0, 0.02]   |

| High risk group (yearly stroke risk > 1.7%) |
|-------------------------------------------|-----------|-------------|-------------|-------------|-------------|
| Base model                               | New model | NRIe 95% CI | NRIn 95% CI | NRI 95% CI |
| CHA₂DS₂-VASc | LR         | 0         | [0, 0]      | 0.11        | [0.1, 0.12] |
| CHA₂DS₂-VASc | ANN        | 0         | [-0.01, 0]  | 0.12        | [0.11, 0.12]| 0.11       | [0.11, 0.12]|
| LR            | ANN        | 0         | [0, 0]      | 0.01        | [0, 0.01]   | 0          | [0, 0.01]   |

Higher scores indicate better performance of the new model when compared to the old model. 

**NRIe**: The net percentage of persons without the event of interest correctly classified downward

**NRIn**: The net percentage of persons with the event of interest correctly classified upward
Development of a regression model and neural network to predict stroke

statistical analysis or differences in study population and data processing or a different
definition of stroke (with or without a TIA).[5, 29]
The main purpose of the CHA2DS2-VASc has been to identify truly low risk patients and
studies have shown warfarin and NOACs to be beneficial for all but the lowest risk catego-
ries (CHA2DS2-VASc = 0).[4] Nevertheless, this study shows that improvements can still be
made in stroke prediction despite the low cutoff for a net benefit. Both logistic regression
and neural networks are capable of making more accurate predictions, especially in low
risk groups, resulting in less overtreatment and thus less bleedings and lower costs.
[29] We found two studies that investigated potential benefits of using other statistical
methods for stroke prevention in AF. These studies showed that current diagnosis and
treatment can be improved by using modeling (in this case, neural networks and Bayesian
statistics).[30, 31] Due to the differences in methodology we cannot directly compare
our results to these studies.
Positive results using modeling methods are also confirmed in another study where the
Framingham risk prediction schemes were compared to more advanced methods.[32]
We attribute the lack of difference in performance between the LR and the ANN to the
fact that many available risk factors are binary, and these data do not have the ‘depth’
from which an ANN can benefit by making complex non-linear relationships between risk
factors. Thus the most likely reason for the enhanced predictive performance of the LR
and ANN is the more accurate weighing of variables that these models are capable of,
as opposed to a ‘sum’ of risk factors. Future studies on the use of these models in stroke
prediction should attempt to include more risk factors as continuous variables.

Variable selection
Studies have mentioned that consistency of AF predictors is suboptimal. Hypertension,
vascular disease, heart failure, diabetes and gender are all mentioned, in separate stud-
ies, as having lower predictive power than core attributes for stroke prediction, namely:
previous stroke and age.[7, 27, 33-35] In this study, age, gender, history of stroke and dia-
betes were consistently selected as predictors of stroke. Hypertension (treated), although
part of both CHADS2 and CHA2DS2-VASc, has been under discussion in other papers. Its
lack of consistency is often attributed to differences in interpretation of blood pressure
measurements or missing data.[7, 35] This could also explain the lack of consistency in
our results. Renal failure was selected in 2 of the 3 models and has been suggested as
a predictor by several other studies which our data appears to support.[5, 33, 36] Smok-
ing, which was selected for both the LR model and the ANN is not often mentioned in
the literature on stroke prediction and warrants further investigation. Lastly, proteinuria,
which has been mentioned as a potential predictor for stroke, was only selected for the
ANN. Again this could be due to interpretation of lab results or missing data. It should
be noted that renal function is usually tested when there is indication to do so, and we
assume there are no renal issues when no tests are documented. This clearly is a strong assumption. The above leads us to conclude that to develop truly accurate and consistent stroke risk prediction models, future research into predictors of stroke should focus on predictors that can be consistently determined from routinely collected data. This is even more relevant as we seek to incorporate biomarkers and patient experience into stroke prediction models.[6] Consistently good external validation of complex prediction scores might be a tall order, and the future may lie in locally optimized prediction models that are based on the local interpretation of routinely collected data.[37]

**Strengths and limitations**

This study used a large cohort of 60595 AF patients, naïve for anticoagulation, to compare different means of predicting stroke risk. It is the first study to compare the CHA2DS2-VASc to complex prediction models for stroke prediction in a large, routinely collected dataset. It therefore represents a real world scenario more accurately, as these prediction methods will mainly be implemented in Electronic Health Record systems. Linkage to HES data ensured only confirmed episodes of stroke were included in the dataset. By assessing reclassification in low and high risk groups we mimicked the clinical situation where only truly low risk patients stand to benefit from being withheld OAC, as opposed to other studies that focused on prediction accuracy rather than net benefit. The ESC guideline formally does not recommend oral anticoagulation for female patients with lone AF (3% of the study population), which might have resulted in a small number CHA2DS2-VASc non-cases to be classified as cases in this study. We did not include bleeding risk assessment into our models due to missing variables, while this is essential in clinical practice. Future studies should investigate whether it is feasible to integrate stroke risk and bleeding risk into a single model and provide physicians with information that displays net benefit of OAC over time for individual patients. The CPRD dataset consisted of only patients who did not already use OAC, which could lead to selection bias, i.e. favoring patients that for some reasons did not get OAC prescribed. Patients were censored at the start of OAC treatment, loss to follow up or death. As antithrombotic treatment improved through the years, patients included in later years may have been followed for a shorter period of time. This could have reduced the number of high risk patients in the dataset in later years. Despite these shortcomings, it should be noted that we currently do not have better methods to investigate novel methods for stroke prediction, as it would be unethical to withhold patients with AF from OAC. The use of random test sets and a separate validation set ensured we limited the chance of over fitting our models and optimized external validity of our results. Although the latter can only truly be established by external validation. Temporal validation was attempted but not feasible due to differences in stroke rate between temporally sampled
training and validation sets, which is likely due to the short follow up period for patients at the end of the dataset. It is possible that other statistical methods are more capable at stroke prediction than LR models or ANN, which requires further investigation.

**Implications for practice**

The findings of this study lead us to conclude that the use of LR or ANN, specifically for decision support, should be investigated. It would be safe to assume that most decision support systems are guideline based, although no studies explicitly investigated this fact. These guidelines contain traditional prediction and classification schemes like the CHA$_2$DS$_2$-VASc, which are developed for use in daily practice and are thus kept simple. In modern day medicine, daily practice revolves around a computer, which is capable of much more than calculating a simple cumulative prediction scheme. This fact will be exaggerated when more complex predictors like biomarkers are added to prediction schemes. By fully utilizing prediction models, prediction of disease and classification of patients can be improved. This has been demonstrated by several studies using regression, Bayesian and neural network methodologies. It is however, paramount to seek external validation before implementing computer based prediction and classification models, as it is in traditional schemes.[38, 39]

**CONCLUSION**

Logistic regression models and neural networks can improve stroke risk predictions and reduce overtreatment when compared to the CHA$_2$DS$_2$-VASc, thus reducing the number of adverse events caused by oral anticoagulation. Future studies will investigate use of these methods in other datasets, but we suggest more studies focus on use of prediction models to classify patients. Furthermore we argue, that to create widely usable, stable (stroke) prediction schemes, research should focus on predictors that can consistently be determined from routinely collected data, or define how these are to be extracted from routinely collected data.

**AUTHOR CONTRIBUTIONS**

RvR provided the data, performed data cleaning and contributed to the manuscript. TvS helped with the initial planning of the study and significantly contributed to the manuscript. DA performed the analyses, and wrote the manuscript. HvW significantly contributed to the manuscript and helped interpreting findings. AA provided input for the statistical analyses and significantly contributed to the manuscript.
REFERENCES

Development of a regression model and neural network to predict stroke


