Universiteit Leiden
The Netherlands

From patient's perspective to predictive modeling: cerebrovascular events and chronic kidney disease
Jong, Y. de

Chapter 5

VALIDATION OF RISK SCORES FOR ISCHEMIC STROKE IN ATRIAL FIBRILLATION ACROSS THE SPECTRUM OF KIDNEY FUNCTION

Ype de Jong
Edouard L. Fu
Merel van Diepen
Marco Trevisan
Karolina Szummer
Friedo W. Dekker
Juan J. Carrero
Gurbey Ocak

To minimize the impact of printing on the environment, the supplementary material has been uploaded. For Chapter 5, the supplement is 33 pages, and can be downloaded using the QR code.

Abstract

Aims: The increasing prevalence of ischemic stroke (IS) can partly be explained by the likewise growing number of patients with chronic kidney disease (CKD). Risk scores have been developed to identify high-risk patients, allowing for personalized anticoagulation therapy. However, predictive performance in CKD is unclear. The aim of this study is to validate six commonly used risk scores for IS in atrial fibrillation (AF) patients across the spectrum of kidney function.

Methods and results: Overall, 36 004 subjects with newly diagnosed AF from SCREAM (Stockholm CREAtinine Measurements), a healthcare utilization cohort of Stockholm residents, were included. Predictive performance of the AFI, $CHADS_2$, Modified $CHADS_2$, CHA_2DS_2-VASc, ATRIA, and GARFIELD-AF risk scores was evaluated across three strata of kidney function: normal kidney function [estimated glomerular filtration rate (eGFR) >60 mL/min/1.73 m^2], mild CKD (eGFR 30–60 mL/min/1.73 m^2), and advanced CKD (eGFR <30 mL/min/1.73 m^2). Predictive performance was assessed by discrimination and calibration. During 1.9 years, 3069 (8.5%) patients suffered an IS. Discrimination was dependent on eGFR: the median c-statistic in normal eGFR was 0.75 (range 0.68–0.78), but decreased to 0.68 (0.58–0.73) and 0.68 (0.55–0.74) for mild and advanced CKD, respectively. Calibration was reasonable and largely independent of eGFR. The Modified $CHADS_2$ score showed good performance across kidney function strata, both for discrimination [c-statistic: 0.78 (95% confidence interval 0.77–0.79), 0.73 (0.71–0.74) and 0.74 (0.69–0.79), respectively] and calibration.

Conclusion: In the most clinically relevant stages of CKD, predictive performance of the majority of risk scores was poor, increasing the risk of misclassification and thus of over- or undertreatment. The Modified $CHADS_2$ score performed good and consistently across all kidney function strata, and should therefore be preferred for risk estimation in AF patients.

5.1 INTRODUCTION

The prevalence of ischemic stroke (IS) is increasing and has become a leading cause of morbidity and mortality worldwide.[1] Chronic kidney disease (CKD) is associated with an increased risk of IS via various mechanisms, both specific to CKD (e.g. accelerated atherosclerotic vascular disease) and general risk factors, such as hypertension, diabetes mellitus, dyslipidemia, and ageing.[2] With an estimated prevalence of 10–15% in the general population, a number that is increasing steadily,[3] CKD may partly explain the high number of strokes.[4,5] Atrial fibrillation (AF), which is considered the main risk factor for IS both in the general population and in CKD patients, is more commonly reported in this fragile population, an observation that may be related to shared risk factors such as age, diabetes, and hypertension.[3,6–8]

Risk scores for IS are essential to weigh the risk of IS vs. the risk of treatment-related bleeding and thus deliver patient-tailored therapy. In patients with CKD, this notion is highly relevant since these patients are at increased risk of treatment-related bleeding as well.[9–11] Typically, most risk scores use clinical parameters (e.g. disease history) in combination with patient-specific characteristics (e.g. age and sex) to compute a risk for IS within a given prediction timeframe. Although widely used risk scores, such as $CHADS_2$ and its updated version CHA_2DS_2-VASc, are endorsed by current guidelines on IS,[8,12–14] their predictive performance in patients with CKD is largely unknown, as these risk scores have been developed in general AF populations.[2,8] For incident dialysis patients, however, external validation studies showed poor predictive performance both for risk scores predicting IS and bleeding.[15,16] Despite these uncertainties, the use of these clinical decision aids is appealing as a seemingly objective tool to standardize the allocation of anticoagulation therapy within CKD care. However, due to the lack of information on the validity of these risk scores in patients beyond the development cohorts of the original studies, their use comes with a risk of misclassification. The aim of the present study is therefore to externally validate multiple commonly used risk scores for IS in a cohort of patients with AF across the spectrum of kidney function.

5.2 METHODS

This study was reported in line with the TRIPOD guideline.[17]

5.2.1 Study population and baseline definition

We used data from the Stockholm CREAtinine Measurements (SCREAM) project, a healthcare utilization cohort from Stockholm, Sweden.[18] SCREAM included all Stockholm

residents aged ≥18 years who had a measurement of serum creatinine from in- or outpatient care between 2006 and 2011. SCREAM includes data from about 1.3 million adults, corresponding to 68% of the population of the region for that period. Information on demographics, disease history, vital status, pharmacy-dispensed medication, and healthcare use was obtained by linking to regional and national administrative databases.[18] All subjects with new-onset AF from January 2007 to December 2012 were selected. New-onset AF was defined as the presence of ICD-10 code I48 in any diagnostic position in primary, outpatient specialist or hospital care, with no I48 code between 1997 (when ICD coding started) and 2007. Baseline was defined as the date of first occurrence of AF. Patients were censored at the end of follow-up (31 December 2012), when they moved outside the Stockholm region or died from other causes than IS. Patients with missing data on creatinine were excluded. Since this study utilized only de-identified data, it was not deemed to require informed consent. The study was approved by the regional ethical review boards and the Swedish National Board of Welfare.

5.2.2 Outcome and predictor definitions

Study outcomes were ascertained via linkage with the government-run National Population Registry, which registers all deaths without loss to follow-up, and the National Patient Registry with diagnosis codes for essentially all (>99%) hospitalizations. The study outcome was defined as hospital admission for IS (ICD-10 codes I63x, I69.3, I69.4, and I69.8 in 1st or 2nd diagnostic position) or IS as main cause of death. Estimated glomerular filtration rate [eGFR; mL/min per $1.73 \, m^2$, calculated with the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula] was calculated using the most recent measurement prior to AF diagnosis (median 0.28 years). Creatinine was measured in plasma, with either an enzymatic or corrected Jaffe method (alkaline picrate reaction); both methods are traceable to isotope dilution mass spectroscopy standards. Creatinine values <25 or >1500 μmol/L were considered outliers and were discarded. Proteinuria (median 0.87 years prior to AF diagnosis) was measured by either urinary albumin-to-creatinine ratio >30, or a urine dipstick (range: negative, 1, 2, and 3; all positive values were regarded as proteinuria). Information on disease history, including previous stroke, previous bleeding, congestive heart failure, cardiovascular disease, hypertension, and diabetes, was obtained using ICD-10 codes (detailed in **Supplementary material chapter 5, Table S1)**. The overall positive predictive value of these diagnoses in the register is about 85–95%.[19] Medication use, including antihypertensive and anti-diabetic drugs, was defined by registered pharmacy dispensations in the 180 days prior to AF diagnosis; for vitamin K antagonist or direct oral anticoagulant usage, dispensations in the 120 days before or up to 60 days after AF diagnosis were evaluated.

5.2.3 Risk scores

Risk scores to be validated were identified from a previous systematic review.[16] Based on availability of predictors, the following risk scores were validated: AFI,[20] CHADS$_2$,[21] Modified CHADS$_2$,[22] CHA$_2$DS$_2$-VASc,[23] ATRIA,[24] and GARFIELD-AF.[25] Scores were validated within the designated timeframe if specified (i.e. the prediction timeframe as specified in the original article), or within the maximum follow-up of the development cohort if no timeframe was specified. We used the same predictor definitions as the original studies where possible. An overview of the included risk scores, the original predictor definitions, and those used in this validation study is presented in the **Supplementary material chapter 5, section '*Risk Scores*'**.

5.2.4 Statistical analysis

The predictive performance of the included risk scores was assessed by their discrimination and calibration abilities, stratified by CKD stage, using the estimated glomerular filtration rate (eGFR) classification of the KDIGO (Kidney Disease: Improving Global Outcomes) criteria. Normal kidney function was defined as KDIGO G1-2 (eGFR >60 mL/min/1.73 m²), mild CKD as KDIGO G3 (eGFR 30–60 mL/min/1.73 m²), and advanced CKD as KDIGO G4-5 (eGFR <30 mL/min/1.73 m²). Discrimination was assessed by the concordance index (c-index or c-statistic), which reflects how well the risk score distinguishes between patients with and without the outcome of interest. The c-statistic lies between 0.5 and 1.0, which equals pure chance and perfect discrimination, respectively. In general, c-statistic <0.7 is considered poor to moderate, 0.8 is considered good, and >0.9 excellent. For logistic risk scores, an area under the receiver operating curve was calculated. For Cox models, Harrell's c-statistic was calculated. Calibration describes the agreement between the predicted and actual probabilities of the outcome. It is typically presented in a calibration plot or calibration in the large (population average observed frequency and average predicted probability). In case of ideal calibration, the slope of the calibration curve would be 1 (i.e. a 45 degree line: predicted probability equals observed probability); for calibration in the large, the average observed and predicted probabilities would be equal. When risk scores presented an event rate instead of cumulative incidence, the cumulative incidence was approximated, as done in a previous study[16] (method detailed in **Supplement material chapter 5, section '*Formulae*'**). To assess the effect of the prediction timeframe (i.e. the time between baseline and when the outcome can occur, e.g. at 1 year for the CHA$_2$DS$_2$-VASc[23] and GARFIELD-AF[25] scores, and 5 years for the Modified CHADS$_2$ score[22]) on the predictive performance, the included risk scores were sequentially validated for different timeframes at monthly intervals, and c-statistics and calibration in the large were plotted over the entire follow-up duration of SCREAM.

This analysis may provide insight into the stability of the performance of risk scores and the dependency on the prediction timeframe. We conducted this analysis since it is not uncommon for clinicians to extrapolate or interpolate the predicted risks over time and we hypothesized that both discrimination and calibration would be highest at the timeframe for which the risk score was developed.

5.2.5 Sensitivity analyses

Three sensitivity analyses were conducted. First, an analysis using a broader, composite outcome definition of IS including transient ischemic attacks (TIAs; ICD-10 codes detailed in **Supplementary material chapter 5, Table S1**). Second, an analysis stratified on anticoagulation use, which included both vitamin K antagonists and direct oral anticoagulants. Lastly, we validated the included risk scores in subgroups with smaller eGFR ranges than the KDIGO stages to further explore the effect of eGFR on score performance. To compare the calibration in the large of the different risk scores, the mean squared error (MSE) of the average predicted and observed probabilities per eGFR cut-off ($n=11$) was calculated. The MSE is the average of the differences between the predicted and observed risks. Lower values indicate a good concordance between these risks, while higher values indicate over- or underprediction, or a combination of both. The methods used to approximate the cumulative incidence and calculate the MSE are further detailed in the **Supplementary material chapter 5, section *'formulae'*.**

5.3 RESULTS

5.3.1 Demographics

Of the 1 372 425 healthcare users in Stockholm included in SCREAM, 39 260 subjects were diagnosed with AF between 2007 and 2011, of which 3256 were excluded because of missing information on eGFR, leaving a total of 36 004 subjects eligible for analysis (**Figure 5.1**). At a median follow-up of 1.88 years, a total of 3069 (8.5%) IS occurred: 1946 (7.4%) of the 26 249 patients with normal kidney function, 1018 (11.8%) of the 8625 patients with mild CKD, and 105 (9.3%) of the 1130 patients with advanced CKD. The baseline characteristics of the included subjects, together with an overview of the study demographics of the validated risk scores, are presented in **Table 5.1.**
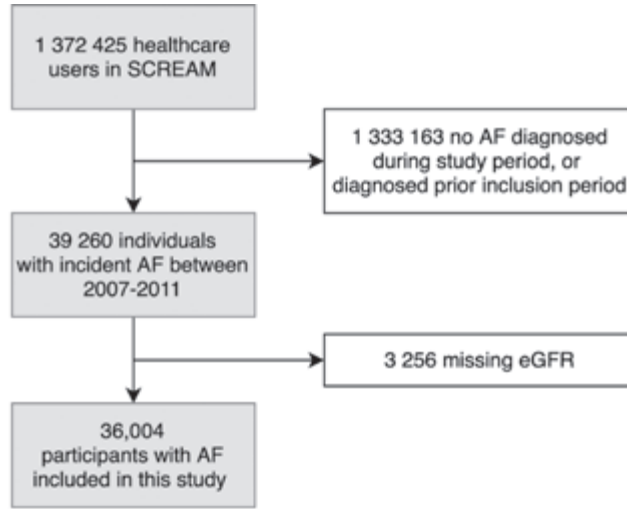
Figure 5.1 Flowchart of patient inclusion in SCREAM. AF: atrial fibrillation.

5.3.2 Discrimination

C-statistics of most risk scores were lower across worsening kidney function categories (**Table 5.2, Figure 5.2**). For the AFI score, c-statistics were 0.68 (95% confidence interval 0.67–0.69) in AF patients with normal kidney function, 0.58 (0.57–0.59) in those with mild CKD and 0.55 (0.51–0.59) in patients with advanced CKD. The c-statistics for the $CHADS_2$ were relatively stable. The Modified $CHADS_2$ score showed the highest and consistent discriminatory abilities in all kidney function groups [0.78 (0.77–0.79), 0.73 (0.71–0.74), and 0.74 (0.69–0.79), respectively]. The CHA_2DS_2-VASc score showed moderate discrimination in AF patients with normal kidney function, but poor discrimination in mild and advanced CKD. The ATRIA risk score showed good discrimination in AF patients with normal kidney function, but moderate in those with mild and advanced CKD, as did the GARFIELD-AF risk score.

Study	Validation cohort	AFI Investigators 1994[17]	Gage 2001[18]	Rietbrock 2008[19]	Lip 2010[20]	Singer 2013[21]	Fox 2017[22]
Name risk score	SCREAM	AFI	CHADS$_2$	Modified-CHADS$_2$	CHA$_2$DS$_2$-VASC	ATRIA‡	GARFIELD-AF
Total participants	36 004	4253	1 733	305 566	1 084	10 927	38 935
Total events (%)	3 069 (8.52)	106 (2.49)	94 (5.42)	19 925 (6.52)	25 (2.31)	685 (6.27)	473 (1.21)
Median follow-up	1.88y	1.2-2.3y‡‡	1.0y	2.46-2.74 y†	1y††	2.4 y	1y††
Age (SD)	74.84 (12.80)	69.4*	81		66	-‡	71.0
Male sex (%)	18 891 (52.5)	2977 (70)*	728 (42)	157 202 (48.6)	642 (59.2)	-‡	21 628 (55.5)
Race (% Caucasian)	-	3930 (92.4)*	-			-	24 157 (62.0)
CKD							
Normal eGFR (%)	26 249 (72.9)	-	-	-	-	-‡	-
Mild CKD (%)	8 625 (24.0)	-	-	-	-	-‡	-
Advanced (%)	1 130 (3.1)	-	-	-	-	-‡	-
Other	-	-	-	-	-	-‡	4038 (12.0)§§
Proteinuria (%)	2 411 (6.7)	-	-	-	-	-‡	-
Diabetes mellitus (%)	7 000 (19.4)	595 (14)*	399 (23)§	-**	187 (17.3)	-‡	8 558 (22.0)
AF (%)	36 004 (100)	4 253 (100)*	1 733 (100)	51 807 (17.0)	1 084 (100)	10 927 (100)	38 935 (100)
Heart failure (%)	9 340 (25.9)	940 (22.1)*	970 (56)§	-**	253 (23.5)	-‡	8 752 (22.5)
Hypertension (%)	21 966 (61.0)	1927 (45.3)*	970 (56)§	-**	729 (67.3)	-‡	30 435 (78.2)
Previous stroke (%)	4 608 (12.8)	264 (6.2)*	433 (25)§	-	45 (4.2)	-‡	3 030 (7.8)
Peripheral vascular disease (%)	3 069 (8.5)	404 (9.5)*	-	-	62 (5.8)	-‡	2212 (5.7)
Ischemic heart disease (%)	9 814 (27.3)	519 (12.2)*	-	-**	412 (38.4)	-‡	-
Previous bleeding (%)	4 969 (13.8)	-	-	-	-	-	1 024 (2.6)
Antithrombotic drug (%)							
VKA usage (%)	14 526 (40.3)	2 113 (49.7)	0 (0)	-	0 (0)	0 (0)	16 491 (42.4)
DOAC usage (%)	217 (0.6)	-	-	-	-	-	8 804 (22.6)
Antiplatelet (%)	23 321 (64.8)	Uncertain number	529 (31)	-	802 (74.0)	-	14 084 (36.2)

Table 5.1 Baseline characteristics of the validation cohort (SCREAM) and the validated risk scores (AFI, CHADS, Modified- CHADS, CHADS$_2$, CHA$_2$DS$_2$-VASC, ATRIA, GARFIELD-AF). CKD categories were defined as normal kidney function as KDIGO G1-2 (eGFR>60 ml/min/1.73m^2), mild CKD as KDIGO G3 (eGFR between 30ml/min/1,73 m^2 and 60ml/min/1,73 m^2) and advanced CKD as KDIGO G4-5 (eGFR below 30ml/min/1.73m^2). *The AFI risk score was presented in four subgroups, with percentages instead of absolute numbers for baseline characteristics. These were estimated by calculating a weighted mean of these percentages. †Median follow-up only given per subgroup. ‡Presented their baseline characteristics per person years. §Presented as rounded percentages; absolute number estimated. **Presented as event rates; no absolute numbers. ††No information was provided on follow-up duration, but patients were censored after reaching this timepoint. ‡‡Range of mean follow-up of the included study populations. §§ Defined as CKD III-V or eGFR <60s. 0.74 (0.69–0.79), respectively]. The CHA$_2$DS$_2$-VASc score showed moderate discrimination in AF patients with normal kidney function, but poor discrimination in mild and advanced CKD. The ATRIA risk score showed good discrimination in AF patients with normal kidney function, but moderate in those with mild and advanced CKD, as did the GARFIELD-AF risk score.
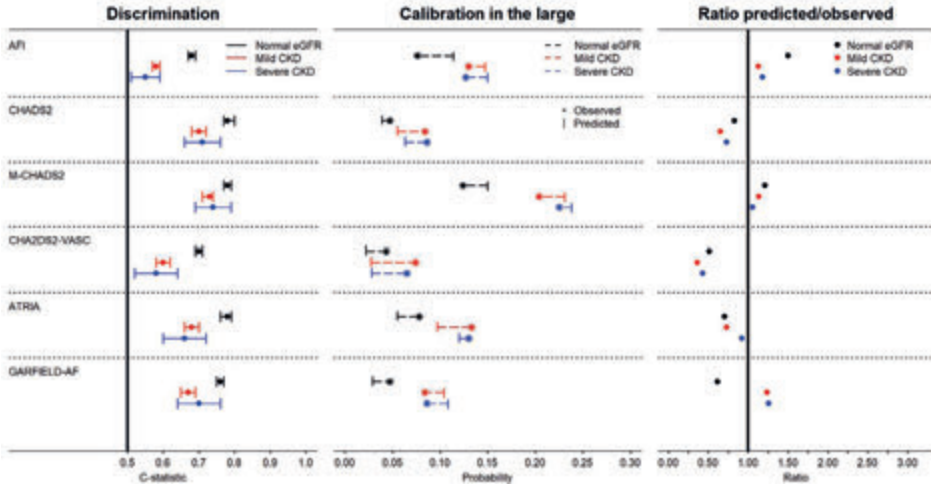
Figure 5.2 Visualization of the predictive performance, stratified by the three eGFR categories. Left: C-statistic (dot) with 95% confidence interval (bar). Middle: calibration in the large, showing the average observed (asterisk) and predicted (bar) probabilities for ischemic stroke. Left: the ratio of the predicted/observed risks – ratio's above one indicate overprediction, ratio's below underprediction. Differences in the observed risks are due to the prediction timeframe of the validated risk scores, and calculation methods (i.e. Cox or logistic).

5.3.3 Calibration

Most risk scores showed modest calibration, largely independent of kidney function (**Figures 5.2 and 5.3; Table 5.2**). For the AFI score, the calibration in the large showed overprediction in all kidney function categories. The $CHADS_2$ score underpredicted risks in the three kidney function categories. The Modified $CHADS_2$ score showed good calibration for the normal eGFR and mild CKD group, but slight overprediction in the advanced CKD group, especially so for the higher-risk patients. The CHA_2DS_2-VASc score underpredicted risks. The ATRIA score underpredicted the risk of IS. Finally, the GARFIELD-AF underpredicted the risk of IS in the normal eGFR category, but overpredicted the risks in mild and advanced CKD. The calibration plots illustrated the inaccuracy of most risk scores for patients with a high risk of IS, regardless of CKD stage, and also underlined the differences in the broadness of the prediction range (i.e. the range of possible predicted risks) (0–0.077 for CHA_2DS_2-VASC to 0.002–0.521 for GARFIELD-AF).

| Study | Risk score characteristics | | | Original C-statistic | Validation | | | | | | | | |
	Design	Outcome	Timeframe (validated)		Normal eGFR			Mild CKD			Advanced CKD		
					C-statistic	Obs	Pred	C-statistic	Obs	Pred	C-statistic	Obs	Pred
AFI Investigators[17] 1994, AFI	Cox	IS, TIA, SE	NS, (2.3y)	-	0.68 (0.67-0.69)	0.076	0.114	0.58 (0.57-0.59)	0.130	0.147	0.55 (0.51-0.59)	0.127	0.150
Gage[18] 2001, CHADS2	Cox	IS, TIA	NS, (1.0y)	0.82	0.78 (0.77-0.80)	0.047	0.039	0.70 (0.68-0.72)	0.084	0.055	0.71 (0.66-0.76)	0.086	0.063
Rietbrock[19] 2008, Modified-CHADS2	Cox	IS, HS	5y; (5y)	0.72	0.78 (0.77-0.79)	0.124	0.150	0.73 (0.71-0.74)	0.204	0.231	0.74 (0.69-0.79)	0.225	0.238
Lips[20] 2010, CHA2DS2-VASC	Logistic	IS, TIA, SE	1y; (1y)	0.61	0.70 (0.69-0.71)	0.043	0.022	0.60 (0.58-0.62)	0.074	0.027	0.58 (0.52-0.64)	0.065	0.028
Singer[21] 2013, ATRIA	Cox	IS, SE	NS, (2.4y)	0.73	0.78 (0.76-0.79)	0.078	0.055	0.68 (0.66-0.70)	0.133	0.097	0.66 (0.60-0.72)	0.130	0.120
Fox[22] 2017, GARFIELD-AF	Cox	IS, TIA, SE	1y; (1y)	0.69	0.76 (0.75-0.77)	0.047	0.029	0.67 (0.65-0.69)	0.084	0.104	0.70 (0.64-0.76)	0.086	0.108

Table 5.2 Overview of the predictive performance of the six included and externally validated studies. Discrimination (C-statistic) and calibration in the large (obs; observed vs pred; predicted) stratified for CKD stages. CKD categories were defined as normal kidney function as KDIGO G1-2 (eGFR>60 ml/min/1.73m²), mild CKD as KDIGO G3 (eGFR between 30ml/min/1.73 m² and 60ml/min/1.73 m²) and advanced CKD as KDIGO G4-5 (eGFR below 30ml/min/1.73m²). Risk scores were validated at the timeframe specified in the article, or if no specification was given, at the maximum follow-up. Abbreviations: RCT: randomized controlled trial; TIA: transient ischemic attack; IS: ischemic stroke; SE: systemic embolus; HS: hemorrhagic stroke; SD: standard deviation.
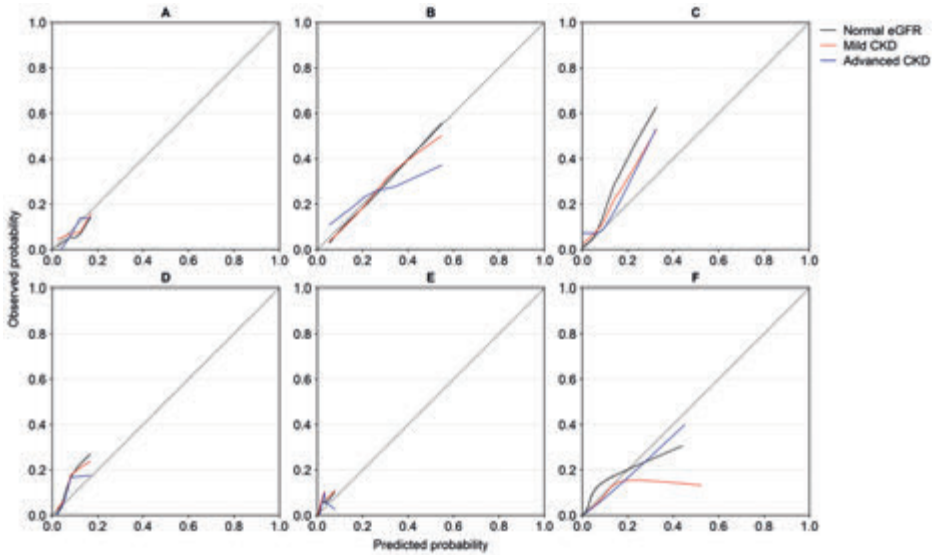
Figure 5.3 Calibration plots showing observed and predicted probabilities for ischemic stroke in patients with CKD and AF in SCREAM. Figure 5.3A: AFI; B: Modified-CHADS$_2$; C ATRIA; D: CHADS$_2$; E: CHA$_2$DS$_2$-VASc; F: GARFIELD-AF

5.3.4 Effect of the prediction timeframe on predictive performance

C-statistics were relatively stable over time, with most risk scores showing only a mild decrease in c-statistic, and subsequently stabilization (**Figure 5.4**, upper panel; stratified for CKD stages see **Supplementary material chapter 5, Figure S9**). For calibration in the large, the optimal prediction timeframe was shorter than in the development studies for CHADS$_2$ (optimal timepoint at 6 months, developed for 12 months), CHA$_2$DS$_2$-VASc (1 and 12 months, respectively), ATRIA (optimal at 17 months, validated at 29 months) and only marginally so for GARFIELD-AF (optimal at 9 months, developed for 12 months), and longer for the AFI (optimal at 49 months, validated at 28 months). The Modified CHADS$_2$ score (developed for 60 months) did not reach an optimal timepoint within 72 months (**Figure 5.4**, lower panel; stratified for CKD stages, **Supplementary material chapter 5, Figure S10**).

5
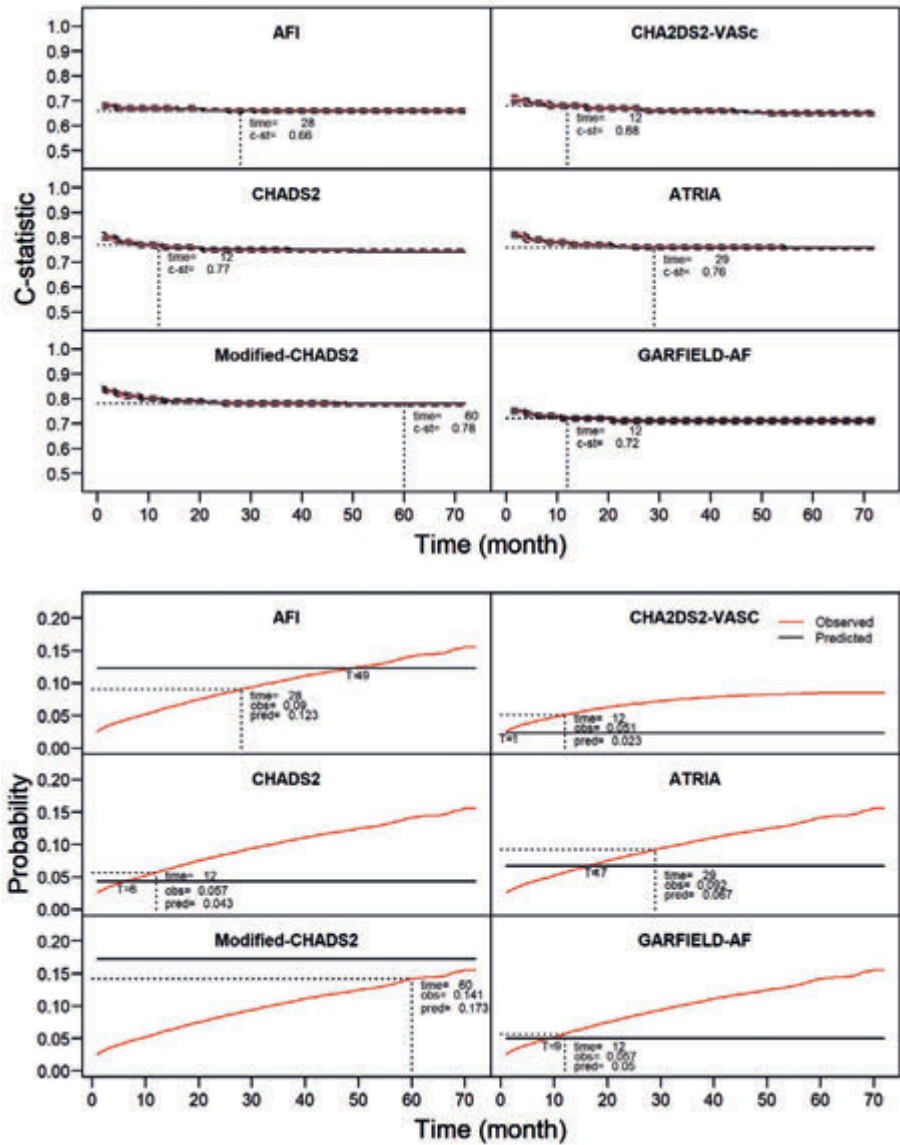
Figure 5.4 Effect of prolonging the prediction timeframe on the predictive performance in patients with AF, not stratified for CKD stage. Above the effect on discrimination (C-statistic with confidence interval); below the effect on calibration in the large. Risk scores were validated 72 times; each time prolonging the prediction timeframe with one month until the maximum follow-up of 72 months was reached. Dotted crosslines indicates the prediction timeframe for which the risk score was developed and the corresponding predictive performance, optimal calibration in the large indicated with T, followed by time in months. Stratification per CKD stage is presented in the **Supplementary material chapter 5, Figures S9-10**.

5.3.5 Sensitivity analyses

For discrimination, when validated for IS and TIA instead of IS only (sensitivity analysis 1, detailed in **Supplementary material chapter 5, Tables S2 and S3; Figure S4**), outcomes were comparable to the main analysis. Stratification by anticoagulation use (sensitivity analysis 2, **Supplementary material chapter 5, Tables S4–S7; Figure S4**) showed similar results, indicating independence from anticoagulation usage, but with broader confidence intervals due to smaller sample sizes. For most risk scores, there was a trend towards poorer discrimination in patients with lower eGFR compared with higher eGFR (sensitivity analysis 3, **Supplementary material chapter 5, Table S8, Figure S6**). The Modified CHADS$_2$ score showed consistently good discriminatory abilities, both in the main analysis and in sensitivity analyses. For calibration, the findings of the main analysis were consistent with the sensitivity analyses 1 and 2 (**Supplementary material chapter 5, Tables S3 and S5; Figures S1- 3, and S5**). The difference between the mean observed and predicted probabilities (calibration in the large) over the eGFR strata (sensitivity analysis 3, **Supplementary material chapter 5, Figure S7**) was stable, as illustrated with the low MSE values, indicating independence of the accuracy of risk scores from eGFR (**Supplementary material chapter 5, Table S9, Figure S8**). Modification of the outcome, using only ICD-10 I63x, showed similar predictive performance, though the number of events decreased to 2572 with corresponding broader confidence intervals for the c-statistics (**Supplementary material chapter 5, Section A**).

5.4 DISCUSSION

In this cohort study of 36 004 patients with AF, we externally validated six commonly used risk scores for IS. Although most risk scores showed moderate to good discrimination in patients with normal kidney function, discrimination was less accurate in moderate and advanced CKD. Calibration was largely independent of kidney function, and most risk scores either over- or underpredicted the risk of IS in one or more CKD categories. The broadness of the prediction range (i.e. the scores' ability to differentiate between low and high risks given the range of possible predicted risks) differed greatly between risk scores. The effect of the prediction timeframe influenced the predictive performance: discrimination showed an initial decrease for the shorter timeframes, but stabilized thereafter, indicating that, with regard to discrimination, risk scores can be used to predict IS on a longer or shorter prediction timeframe than designed in the original studies. For calibration, the optimal prediction timepoint differed substantially with the timepoint in the original study of most risk scores. Our results support the use of the Modified CHADS$_2$ score[22] in clinical practice as it showed good and consistent discrimination and calibration in all three kidney function categories.

Given the increasing prevalence of CKD, and the frequent use of risk scores for IS in the care of patients with CKD, there is surprisingly little information on the predictive performance in this high-risk population. Except for GARFIELD-AF and ATRIA, none of the validated risk scores included patients with CKD in their development cohorts, or CKD-specific predictors[26] in their risk score. Furthermore, external validation—the cornerstone for assessment of predictive performance in 'real-world' patients—of these risk scores is essential, but seldom performed. So far only a few studies included patients with CKD in their validation cohorts, with conflicting results: one large study on 14 264 patients with AF and eGFR >30 mL/min validated both the CHA_2DS_2-VASc and $CHADS_2$ scores showing poor discrimination (c-statistic of 0.578 and 0.575, respectively), but did not present information on calibration.[27] In another study on 307 351 patients with AF, these two risk scores performed considerably better and more in line with our results (c-statistics of 0.71 and 0.72, respectively). However, again no information on calibration was reported.[28] Finally, several studies with substantial smaller sample sizes evaluated the same risk scores, yielding comparable results, but as with the previous studies, none calculated the agreement between observed and predicted risks.[29–32] Yet, from a clinical point of view, it could be argued that this calibration, which indicates the precision of the predicted absolute risks, is clinically more important than discrimination in the setting of weighing the risks of IS and severe bleeding due to anticoagulation. This is especially relevant for patients with AF and CKD, as both the risks of IS and severe bleeding are increased.[9–11] For the clinician facing such a patient, using a risk score may seem an objective method to decide on anticoagulation therapy. However, as our study demonstrates, most of the validated risk scores for IS in this clinically relevant population either substantially over- or underpredict this risk. Although the Modified $CHADS_2$ score showed reasonable performance and would currently be the preferred risk estimation tool for patients with AF and CKD, ideally new risk scores should be developed and validated in this high-risk population. Prediction of bleeding risk appears to be equally influenced by kidney function, though data are only available for patients on dialysis.[12] This effect on predictive performance is not without consequence. Underprediction of IS risk, when weighed with bleeding risk, will result in less patients being treated with anticoagulation and consequently, an increased IS incidence, while overprediction will result in overtreatment and increased bleeding incidence. Regardless, most clinical guidelines on IS prevention recommend using the CHA_2DS_2-VASc score[8,12,13]—which showed poor predictive performance in patients with and without CKD alike. Finally, the AFI, Modified $CHADS_2$, ATRIA, and GARFIELD-AF risk scores have not been validated in patients with CKD.

Predictive performance decreased in the more clinically relevant groups of mild and advanced CKD, especially so for discrimination. Two mechanisms may have contributed to this. First, most risk scores were developed in general AF patients, and most of these studies did not include patients with CKD in their development cohort. When validating these risk scores that were developed in such heterogeneous populations in a more homogeneous population, such as patients with CKD, predictive performance—and especially discrimination—may drop.[33] While the included predictors may predict well in general AF cohorts, other more CKD-specific predictors of IS may better discriminate in this relatively homogeneous population. These include for example eGFR (which was used in ATRIA and GARFIELD-AF), proteinuria (used in ATRIA), primary kidney disease, presence of atherosclerotic vascular disease,[2] or various biomarkers (e.g. myeloperoxidase[34] or fibroblast growth factor-23,[35] amongst others[36]). Second, although we expected a comparable drop in c-statistics for the even more homogeneous patients with advanced CKD, the c-statistics of these groups were roughly equal. This may have been due to chance however: the absolute number of events in the advanced CKD group was smaller, and the level of precision consequently lower. Finally, while we ensured conformity between the predictor definitions of the original studies and our validation cohort, we deliberately validated these risk scores for the same outcome definition of IS. Indeed, most studies were developed to predict the probability of a composite outcome (e.g. CHA_2DS_2-VASc predicts a composite outcome of IS, TIA, peripheral and pulmonary embolisms). Although predictive performance might improve from validating each risk score for their specific outcome, comparability of these risk scores would then become impossible, especially when the composite outcome includes counterintuitive components, such as IS and hemorrhagic stroke. Another reason for using this outcome definition is the clinical usage: these risk scores are usually used for prediction of IS alone in the clinical setting. To test this effect, we included TIA as a composite outcome in a sensitivity analysis, which was included in most risk scores as part of the outcome. As this did not alter the results, we do not expect the effect of this outcome definition to be substantial.

5.4.1 Strengths and limitations

Our study has several strengths, but also limitations. The main strength is our large and well-defined source population, which allowed for a head-to-head comparison of multiple risk scores in well-characterized participants. Our study also provides information on calibration - the agreement between the predicted and observed risks - information that is essential for weighing IS and bleeding risk. Consistently with previous studies,[4,5] patients with more severe CKD stages in our study had a markedly increased risk of stroke.

A first limitation of our study is the large proportion of anticoagulation users in our population. Stratification for anticoagulation users and non-users yielded similar results, although discrimination was slightly poorer in anticoagulation users. Second, the cut-offs for the CKD groups might have influenced the predictive performance. Although of clinical relevance and in line with the KDIGO classification, we aimed to further explore the correlation between discrimination and kidney function. When stratified in smaller groups than the three kidney function groups, it was again shown that for most risk scores discrimination decreased with worsening of kidney function, while calibration remained relatively stable. Third, Swedish regulations do not allow the recording of ethnicity in registers, and we assumed our population to be primarily Caucasian.[37] Disparities in IS risk may be explained by ethnicity,[38] for example, blacks have a two-fold increased risk of stroke compared with non-Hispanic white adults,[39] and the predictive performance of two different scores (QRISK2 and Framingham scores) was indeed influenced by ethnicity.[40] In line with recent debates on the adequacy of the correction factor for African American ethnicity[41–43] in eGFR calculation, extrapolation of our results to other ethnicities should be done with caution. Fourth, because the prediction timeframes differed for the validated risk scores, we were unable to formally compare the predictive performance. Fifth, the use of routinely collected laboratory data may be a source of bias: for example proteinuria, a predictor used for one study (ATRIA), is not routinely measured, and measurements are performed in persons at risk. Finally, in daily clinical practice, it is not uncommon to categorize or dichotomize risk scores (e.g. CHA_2DS_2-VASc is often categorized in zero points, one point, or greater than one). Dichotomization results in loss of information[44] and our sensitivity analysis showed poor performance when validated in commonly used categories. Furthermore, most risk scores were updated many times after publication. We decided to validate the scores as intended by the authors of the original scores, instead of choosing one of the many updates or categorizations, although this may not represent the clinical use of these risk scores.

5.4.2 Implications and conclusion

Our study demonstrated moderate to poor predictive performance of various risk scores for IS in patients with AF and CKD and emphasizes how difficult this prediction is, underlining the statistical work that needs to be done in the field. For most risk scores, discriminatory abilities decreased in clinically relevant patients with mild and advanced CKD. However, calibration, which is essential for weighing the risk of IS and treatment-related bleeding, was less dependent on kidney function but still most risk scores either over- or underpredicted IS risk, or a combination of both. Prediction of IS

risk should be accurate and weighed against the risk of treatment-related bleeding. To this aim, either new scores incorporating CKD-specific predictors should be developed, or alternatively, existing and externally validated scores should be combined to increase predictive performance in this clinically relevant population, for example using ensemble modelling. As most risk scores used different prediction timeframes, this was unfeasible in our study. By conducting a head-to-head comparison of multiple scores , this study provides the clinician with information on which risk score perform well for different prediction timeframes. The Modified CHADS$_2$ score showed the best and most consistent predictive performance in all CKD stages and we suggest it is the preferred risk score to apply in clinical practice. These findings can inform the choice of risk scores in clinical practice, particularly in patients with mild to severe forms of CKD, which have not always been considered when these scores were developed.

5

5.5 REFERENCES

1.   Feigin VL , Norrving B , Mensah GA. Global burden of stroke. Circ Res 2017;120:439–448.

2.   Kumar S , Lim E , Covic A , Verhamme P , Gale CP , Camm AJ , Goldsmith D. Anticoagulation in concomitant chronic kidney disease and atrial fibrillation: JACC review topic of the week. J Am Coll Cardiol 2019;74:2204–2215.

3.   Turakhia MP , Blankestijn PJ , Carrero J-J , Clase CM , Deo R , Herzog CA , Kasner SE , Passman RS , Pecoits-Filho R , Reinecke H , Shroff GR , Zareba W , Cheung M , Wheeler DC , Winkelmayer WC , Wanner C , Amann K , Banerjee D , Bansal N , Boriani G , Bunch J , Chan CT , Charytan DM , Conen D , Friedman AN , Genovesi S , Holden RM , House AA , Jadoul M , Jardine AG , Johnson DW , Jun M , Labriola L , Mark PB , McCullough PA , Nolin TD , Potpara TS , Pun PH , Ribeiro ALP , Rossignol P , Shen JI , Sood MM , Tsukamoto Y , Wang AY-M , Weir MR , Wetmore JB , Wranicz JK , Yamasaki H , Conference Participants. Chronic kidney disease and arrhythmias: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) controversies conference. Eur Heart J 2018;39:2314–2325.

4.   Cherng Y-G , Lin C-S , Shih C-C , Hsu Y-H , Yeh C-C , Hu C-J , Chen T-L , Liao C-C. Risk and outcomes in patients with chronic kidney disease or end-stage renal disease: two nationwide studies. PLoS One 2018;13:e0191155.

5.   Masson P , Webster AC , Hong M , Turner R , Lindley RI , Craig JC. Chronic kidney disease and the risk of stroke: a systematic review and meta-analysis. Nephrol Dial Transplant 2015;30:1162–1169.

6.   Alonso A , Lopez FL , Matsushita K , Loehr LR , Agarwal SK , Chen LY , Soliman EZ , Astor BC , Coresh J. Chronic kidney disease is associated with the incidence of atrial fibrillation: the Atherosclerosis Risk in Communities (ARIC) study. Circulation 2011;123:2946–2953.

7.   Carrero JJ , Trevisan M , Sood MM , Bárány P , Xu H , Evans M , Friberg L , Szummer K. Incident atrial fibrillation and the risk of stroke in adults with chronic kidney disease: the Stockholm CREAtinine Measurements (SCREAM) project. Clin J Am Soc Nephrol 2018;13:1314–1320.

8.   Kirchhof P , Benussi S , Kotecha D , Ahlsson A , Atar D , Casadei B , Castella M , Diener H-C , Heidbuchel H , Hendriks J , Hindricks G , Manolis AS , Oldgren J , Popescu BA , Schotten U , Van Putte B , Vardas P , Agewall S , Camm J , Baron Esquivias G , Budts W , Carerj S , Casselman F , Coca A , De Caterina R , Deftereos S , Dobrev D , Ferro JM , Filippatos G , Fitzsimons D , Gorenek B , Guenoun M , Hohnloser SH , Kolh P , Lip GYH , Manolis A , McMurray J , Ponikowski P , Rosenhek R , Ruschitzka F , Savelieva I , Sharma S , Suwalski P , Tamargo JL , Taylor CJ , Van Gelder IC , Voors AA , Windecker S , Zamorano JL , Zeppenfeld K. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Eur Heart J 2016;37:2893–2962.

9.   Bonde AN , Lip GY , Kamper AL , Hansen PR , Lamberts M , Hommel K , Hansen ML , Gislason GH , Torp-Pedersen C , Olesen JB. Net clinical benefit of antithrombotic therapy in patients with atrial fibrillation and chronic kidney disease: a nationwide observational cohort study. J Am Coll Cardiol 2014;64:2471–2482.

10. Olesen JB , Lip GY , Kamper AL , Hommel K , Køber L , Lane DA , Lindhardsen J , Gislason GH , Torp-Pedersen C. Stroke and bleeding in atrial fibrillation with chronic kidney disease. N Engl J Med 2012;367:625–635.

11. Shah M , Avgil Tsadok M , Jackevicius CA , Essebag V , Eisenberg MJ , Rahme E , Humphries KH , Tu JV , Behlouli H , Guo H , Pilote L. Warfarin use and the risk for stroke and bleeding in patients with atrial fibrillation undergoing dialysis. Circulation 2014;129:1196–1203.

12. Camm AJ , Lip GYH , De Caterina R , Savelieva I , Atar D , Hohnloser SH , Hindricks G , Kirchhof P , Bax JJ , Baumgartner H , Ceconi C , Dean V , Deaton C , Fagard R , Funck-Brentano C , Hasdai D , Hoes A , Kirchhof P , Knuuti J , Kolh P , McDonagh T , Moulin C , Popescu BA , Reiner Ž , Sechtem U , Sirnes PA , Tendera M , Torbicki A , Vahanian A , Windecker S , Vardas P , Al-Attar N , Alfieri O , Angelini A , Blömstrom-Lundqvist C , Colonna P , De Sutter J , Ernst S , Goette A , Gorenek B , Hatala R , Heidbüchel H , Heldal M , Kristensen SD , Kolh P , Le Heuzey J-Y , Mavrakis H , Mont L , Filardi PP , Ponikowski P , Prendergast B , Rutten FH , Schotten U , Van Gelder IC , Verheugt FWA , Authors/Task Force Members. 2012 focused update of the ESC Guidelines for the management of atrial fibrillation: an update of the 2010 ESC Guidelines for the management of atrial fibrillation—developed with the special contribution of the European Heart Rhythm Association. Europace 2012;14:1385–1413.

13. Kernan WN , Ovbiagele B , Black HR , Bravata DM , Chimowitz MI , Ezekowitz MD , Fang MC , Fisher M , Furie KL , Heck DV , Johnston SC , Kasner SE , Kittner SJ , Mitchell PH , Rich MW , Richardson D , Schwamm LH , Wilson JA. Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack. Stroke 2014;45:2160–2236.

14. Hindricks G , Potpara T , Dagres N , Arbelo E , Bax JJ , Blomström-Lundqvist C , Boriani G , Castella M , Dan GA , Dilaveris PE , Fauchier L , Filippatos G , Kalman JM , La Meir M , Lane DA , Lebeau JP , Lettino M , Lip GYH , Pinto FJ , Thomas GN , Valgimigli M , Van Gelder IC , Van Putte BP , Watkins CL ; ESC Scientific Document Group. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association of Cardio-Thoracic Surgery (EACTS). Eur Heart J 2021;42:373–498.

15. Ocak G , Ramspek C , Rookmaaker MB , Blankestijn PJ , Verhaar MC , Bos WJW , Dekker FW , van Diepen M. Performance of bleeding risk scores in dialysis patients. Nephrol Dial Transplant 2019;34:1223–1231.

16. de Jong Y , Ramspek CL , van der Endt VHW , Rookmaaker MB , Blankestijn PJ , Vernooij RWM , Verhaar MC , Bos WJW , Dekker FW , Ocak G , van Diepen M. A systematic review and external validation of stroke prediction models demonstrates poor performance in dialysis patients. J Clin Epidemiol 2020;123:69–79.

17. Collins GS , Reitsma JB , Altman DG , Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594.

18. Runesson B , Gasparini A , Qureshi AR , Norin O , Evans M , Barany P , Wettermark B , Elinder CG , Carrero JJ. The Stockholm CREAtinine Measurements (SCREAM) project: protocol overview and regional representativeness. Clin Kidney J 2016;9:119–127.

19. Ludvigsson JF , Andersson E , Ekbom A , Feychting M , Kim J-L , Reuterwall C , Heurgren M , Olausson PO. External review and validation of the Swedish National Inpatient Register. BMC Public Health 2011;11:450.

20. Atrial Fibrillation Investigators. Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation. Analysis of pooled data from five randomized controlled trials. Arch Intern Med 1994;154:1449–1457.

21. Gage BF , Waterman AD , Shannon W , Boechler M , Rich MW , Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA 2001;285:2864–2870.

22. Rietbrock S , Heeley E , Plumb J , van Staa T. Chronic atrial fibrillation: incidence, prevalence, and prediction of stroke using the Congestive heart failure, Hypertension, Age >75, Diabetes mellitus, and prior Stroke or transient ischemic attack (CHADS2) risk stratification scheme. Am Heart J 2008;156:57–64.

5

23. Lip GY , Nieuwlaat R , Pisters R , Lane DA , Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010;137:263–272.

24. Singer DE , Chang Y , Borowsky LH , Fang MC , Pomernacki NK , Udaltsova N , Reynolds K , Go AS. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. J Am Heart Assoc 2013;2:e000250.

25. Fox KAA , Lucas JE , Pieper KS , Bassand JP , Camm AJ , Fitzmaurice DA , Goldhaber SZ , Goto S , Haas S , Hacke W , Kayani G , Oto A , Mantovani LG , Misselwitz F , Piccini JP , Turpie AGG , Verheugt FWA , Kakkar AK. 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. BMJ Open 2017;7:e017157.

26. Ramspek CL , de Jong Y , Dekker FW , van Diepen M. Towards the best kidney failure prediction tool: a systematic review and selection aid. Nephrol Dial Transplant 2020;35:1527–1538.

27. Ravera M , Bussalino E , Fusaro M , Di Lullo L , Aucella F , Paoletti E. Systematic DOACs oral anticoagulation in patients with atrial fibrillation and chronic kidney disease: the nephrologist's perspective. J Nephrol 2020;33:483–495.

28. Friberg L , Benson L , Lip GY. Balancing stroke and bleeding risks in patients with atrial fibrillation and renal failure: the Swedish Atrial Fibrillation Cohort study. Eur Heart J 2015;36:297–306.

29. Cedeño Mora S , Goicoechea M , Torres E , Verdalles Ú , Pérez de José A , Verde E , García de Vinuesa S , Luño J. Cardiovascular risk prediction in chronic kidney disease patients. Nefrologia 2017;37:293–300.

30. Roldán V , Marín F , Manzano-Fernandez S , Fernández H , Gallego P , Valdés M , Vicente V , Lip GY. Does chronic kidney disease improve the predictive value of the CHADS2 and CHA2DS2-VASc stroke stratification risk scores for atrial fibrillation? Thromb Haemost 2013;109:956–960.

31. Bautista J , Bella A , Chaudhari A , Pekler G , Sapra KJ , Carbajal R , Baumstein D. Advanced chronic kidney disease in non-valvular atrial fibrillation: extending the utility of R2CHADS2 to patients with advanced renal failure. Clin Kidney J 2015;8:226–231.

32. Banerjee A , Fauchier L , Vourc'h P , Andres CR , Taillandier S , Halimi JM , Lip GY. Renal impairment and ischemic stroke risk assessment in patients with atrial fibrillation: the Loire Valley Atrial Fibrillation Project. J Am Coll Cardiol 2013;61:2079–2087.

33. Vergouwe Y , Moons KG , Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010;172:971–980.

34. Correa S , Pena-Esparragoza JK , Scovner KM , Waikar SS , Mc Causland FR. Myeloperoxidase and the risk of CKD progression, cardiovascular disease, and death in the Chronic Renal Insufficiency Cohort (CRIC) study. Am J Kidney Dis 2020;76:32–41.

35. Scialla JJ , Xie H , Rahman M , Anderson AH , Isakova T , Ojo A , Zhang X , Nessel L , Hamano T , Grunwald JE , Raj DS , Yang W , He J , Lash JP , Go AS , Kusek JW , Feldman H , Wolf M , the Chronic Renal Insufficiency Cohort (CRIC) Study Investigators. Fibroblast growth factor-23 and cardiovascular events in CKD. J Am Soc Nephrol 2014;25:349–360.

36. Stenvinkel P , Carrero JJ , Axelsson J , Lindholm B , Heimbürger O , Massy Z. Emerging biomarkers for evaluating cardiovascular risk in the chronic kidney disease patient: how do new pieces fit into the uremic puzzle? Clin J Am Soc Nephrol 2008;3:505–521.

37. Foreign-born persons in Sweden by country of birth, age and sex. Year 2000-2015: Statistics Sweden; 2016. https://www.scb.se/en/ (16 December 2020).

38. Howard VJ , Kleindorfer DO , Judd SE , McClure LA , Safford MM , Rhodes JD , Cushman M , Moy CS , Soliman EZ , Kissela BM , Howard G. Disparities in stroke incidence contributing to disparities in stroke mortality. Ann Neurol 2011;69:619–627.

39. Roger VL , Go AS , Lloyd-Jones DM , Benjamin EJ , Berry JD , Borden WB , Bravata DM , Dai S , Ford ES , Fox CS , Fullerton HJ , Gillespie C , Hailpern SM , Heit JA , Howard VJ , Kissela BM , Kittner SJ , Lackland DT , Lichtman JH , Lisabeth LD , Makuc DM , Marcus GM , Marelli A , Matchar DB , Moy CS , Mozaffarian D , Mussolino ME , Nichol G , Paynter NP , Soliman EZ , Sorlie PD , Sotoodehnia N , Turan TN , Virani SS , Wong ND , Woo D , Turner MB ; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2012 update: a report from the American Heart Association. Circulation 2012;125:e2–e220.

40. Tillin T , Hughes AD , Whincup P , Mayet J , Sattar N , McKeigue PM , Chaturvedi N , Group SS , On behalf of the SABRE Study Group. Ethnicity and prediction of cardiovascular disease: performance of QRISK2 and Framingham scores in a U.K. tri-ethnic prospective cohort study (SABRE–Southall And Brent REvisited). Heart 2014;100:60–67.

41. Diao JA , Wu GJ , Taylor HA , Tucker JK , Powe NR , Kohane IS , Manrai AK. Clinical implications of removing race from estimates of kidney function. JAMA 2021;325:184–186.

42. Powe NR. Black kidney function matters: use or misuse of race? JAMA 2020;324:737–738.

43. Eneanya ND , Yang W , Reese PP. Reconsidering the consequences of using race to estimate kidney function. JAMA 2019;322:113–114.

44. Moons KGM , Wolff RF , Riley RD , Whiting PF , Westwood M , Collins GS , Reitsma JB , Kleijnen J , Mallett S. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1–W33.

5