Prediction of rates of thromboembolic and major bleeding outcomes with dabigatran or warfarin among patients with atrial fibrillation: new initiator cohort study

Shirley V Wang, Jessica M Franklin, Robert J Glynn, Sebastian Schneeweiss, Wesley Eddings, Joshua J Gagne

Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont St, Boston, MA 02120, USA

Correspondence to: S V Wang swang27@partners.org

Additional material is published online only. To view please visit the journal online.

ABSTRACT

OBJECTIVES

To compare stratified event rates from randomized controlled trials with predicted event rates from models developed in observational data, and assess their ability to accurately capture observed rates of thromboembolism and major bleeding for patients treated with dabigatran or warfarin as part of routine care.

DESIGN

New initiator cohort study.

SETTING

Data from United Health (October 2009 to June 2013), a commercial healthcare claims database in the United States.

PARTICIPANTS

21 934 adults with atrial fibrillation initiating dabigatran (150 mg dose only) or warfarin treatment as part of routine care.

MAIN OUTCOME MEASURES

Predicted annual rates of thromboembolism or major bleeding, based on estimates from randomized controlled trials, models developed in routine care patients, and baseline risk scores (CHADS$_2$, CHA$_2$DS$_2$-VASc, and HAS-BLED). Thromboembolism was a composite outcome, including primary inpatient diagnosis codes for ischemic or ill defined stroke, transient ischemic attack, pulmonary embolism, deep vein thrombosis, and systemic embolism. Major bleeding was a composite outcome including codes occurring in an inpatient setting for hemorrhagic stroke; major upper, lower, or unspecified gastrointestinal bleed; and major urogenital or other bleed.

RESULTS

6516 (30%) and 15 418 (70%) of patients initiated dabigatran and warfarin, respectively. Annual event rates per 100 patients were 1.7 for thromboembolism and 4.6 for major bleeding. For thromboembolism, calibration of estimates from randomized controlled trials was similar to calibration for model based predictions; however, trial estimates for major bleeding consistently underestimated the rate of bleeding among patients in routine care. Underestimation of bleeding rates was particularly pronounced in warfarin initiators with high HAS-BLED scores, where event rates were underestimated by up to 4.0 per 100 patient years. Harrell's c indices for discrimination for thromboembolism or major bleeding in dabigatran and warfarin initiators ranged between 0.59 and 0.66 for randomized controlled trial predictions, and between 0.52 and 0.70 for cross validated model based predictions.

CONCLUSION

Estimated rates of thromboembolism under dabigatran or warfarin treatment in randomized controlled trials were close to observed rates in routine care patients. However, rates of major bleeding were underestimated. Models developed in routine care patients can provide accurate, tailored estimates of risk and benefit under alternative treatment to enhance patient centered care.

WHAT IS ALREADY KNOWN ON THIS TOPIC

CHADS$_2$ and HAS-BLED are validated risk scores that estimate baseline risk of thromboembolism or major bleeding, respectively, in patients with atrial fibrillation to guide whether to initiate anticoagulation therapy; but these scores do not indicate which anticoagulant drug to use

Recent randomized controlled trials provide estimates of risk for thromboembolism and major bleeding under treatment with dabigatran or warfarin within low, medium, and high baseline risk categories

Trials have strict inclusion and exclusion criteria that can result in distributions of patient characteristics that are different from those observed in patients treated as part of routine care

WHAT THIS STUDY ADDS

Models developed and validated using observational data performed as well as randomized controlled trials at predicting thromboembolism under treatment with dabigatran or warfarin in patients treated as part of routine care

These models performed better than trials at predicting major bleeding under treatment with dabigatran or warfarin in patients treated as part of routine care

Trials could underestimate the rate of major bleeding in routine care patients at the highest baseline risk for bleeding by up to 4.0 per 100 patient years

Introduction

Widely recognized risk scores such as the CHADS$_2$ score or CHA$_2$DS$_2$-VASc score for stroke and thromboembolism and HAS-BLED score for major bleeding are used to estimate individual risk of these outcomes among patients with atrial fibrillation.[1-4] The CHADS$_2$ score accounts for congestive heart failure, hypertension, age ($\geq$75 years), diabetes mellitus, and previous stroke/transient ischaemic attack/thromboembolism (doubled risk weight), whereas the CHA$_2$DS$_2$-VASc score also accounts for vascular disease, age 65-74 years, and sex category (age$\geq$75 years and previous stroke carry doubled risk weight). The HAS-BLED score accounts for (age>65 years, hypertension, abnormal renal and liver function, prior stroke, bleeding history (or predisposition), drugs predisposing

to bleed, alcohol use disorders, and labile international normalized ratio.

While these risk scores provide baseline estimates of risk to guide whether to treat with an oral anticoagulant, they do not provide insight into optimal treatment selection when treatment is clearly indicated. The efficacy of oral anticoagulants in preventing thromboembolism in patients with atrial fibrillation is well known, with many studies reporting over 60%[5-8] reduction in stroke for patients with non-valvular atrial fibrillation. Also known is the potential for anticoagulants to increase the risk of major bleeding.[6-8]

The updated 2012 European Society of Cardiology guidelines[8] and the National Institute for Health and Care Excellence guidelines[9] recommend that patients with at least one risk factor included in the CHA_2DS_2-VASc score be considered for anticoagulation therapy (excluding female patients under age 65 years with lone atrial fibrillation), and that anticoagulation should be offered to patients with atrial fibrillation and a CHA_2DS_2-VASc score of 2 or more.[8 9] However, which anticoagulant to offer remains debated.[8-10] When a clinician and patient decide to initiate anticoagulation therapy, accurate predictions of benefit (prevention of thromboembolism) or harm (increased risk of major bleeding) with different oral anticoagulation agents can help guide treatment choice.

The Randomized Evaluation of Long-Term Anticoagulation Therapy (RE-LY) trial compared the effectiveness of dabigatran, a novel oral anticoagulation agent, to warfarin. The trial found that, on average, the risk of stroke was 34% lower in participants randomized to receive the 150 mg twice daily dose of dabigatran than in those randomized to receive warfarin.[3] Reductions in risk between 20-40% were found in observational studies in large healthcare databases comparing the effectiveness of dabigatran to warfarin.[11 12] However, effectiveness of oral anticoagulation therapy could vary with baseline risk of thromboembolic or major bleeding outcomes. To provide evidence based, patient centered care that is tailored to patients with different risk factors, additional evidence beyond an average treatment effect for compared therapeutic options is necessary.

In recent trials comparing novel oral anticoagulant agents with warfarin, rates of thromboembolism or major bleeding under treatment with alternative oral anticoagulation agents were provided for patients at baseline risk of these events. Baseline risk of these events was categorized as low, medium, or high, based on the $CHADS_2$ and HAS-BLED risk scores (low=0-1, medium=2, high risk≥3).[1-4] However, estimates based on the highly selected populations eligible to participate in trials might not reflect the baseline risk among populations actually treated in practice.

The objective of the study was to compare stratified event rates from randomized controlled trials with predicted event rates from models developed in observational data, and assess their ability to accurately capture observed rates of thromboembolism and major bleeding under treatment with dabigatran or warfarin as part of routine care. The purpose of obtaining accurate predicted event rates in patients at varying baseline outcome risk was to answer the clinical question: what are the expected rates of thromboembolism and major bleeding for a given patient if that patient were to initiate treatment with dabigatran or warfarin?

Methods

Patients treated as part of routine care

We identified a population of patients initiating dabigatran or warfarin between October 2009 and June 2013 from Optum Life Sciences' longitudinal database of commercial healthcare claims. This database comprised of claims data on over 14 million enrollees in United Health, a large private insurer in the United States.[13] The database includes demographic and enrollment information, diagnoses and procedures from inpatient and outpatient healthcare claims as well as dispensed outpatient medications.

The study cohort comprised patients initiating either warfarin or dabigatran treatment (150 mg dose only). The index date for new initiation was defined as the first claim for dabigatran or warfarin occurring after at least 365 days of enrollment with United Health without a filled claim for an oral anticoagulation agent (dabigatran, warfarin, rivaroxaban or apixaban). We restricted the cohort to include patients over the age of 18 years with at least one recorded diagnosis of atrial fibrillation according to ICD-9 (international classification of diseases, 9th revision). We further restricted the cohort to the overlapping region of the propensity score.[14] A flow diagram[15] for construction of the cohort and additional details on propensity score estimation and trimming are available in the web appendix.

Outcomes, follow-up, and covariates

For each patient, follow-up for outcomes began the day after treatment initiation. Patients were censored at occurrence of the outcome of interest, treatment discontinuation (14 day grace period from end of days' supply), death, disenrollment, prescription dispensation of another anticoagulant, or end of available data. Thromboembolism was a composite outcome, including primary inpatient diagnosis codes for ischemic or ill defined stroke, transient ischemic attack, pulmonary embolism, deep vein thrombosis, and systemic embolism. Major bleeding was a composite outcome including codes occurring in an inpatient setting for hemorrhagic stroke; major upper, lower, or unspecified gastrointestinal bleed; and major urogenital or other bleed. Outcomes and risk factors were defined based on ICD-9 diagnosis, ICD-9 procedure, or Current Procedural Terminology codes. Code lists and citations to validation studies are available in the web appendix. As a supplementary analysis, we evaluated a thromboembolism outcome that did not include transient ischemic attack, pulmonary embolism, or deep vein thrombosis.

We assessed the presence or absence of known risk factors included in the $CHADS_2$ and HAS-BLED scores

using claims occurring on the index date or during the year before. Although recent randomized controlled trials used the CHADS$_2$ score to categorize participants into low, medium, and high risk categories, the CHADS$_2$ score was updated after the trial protocols were conceived. The updated risk score, CHA$_2$DS$_2$-VASc, includes all risk factors in CHADS$_2$ as well as additional risk factors for predicting thromboembolism and is regarded to be a better score for stratifying risk of thromboembolism in patients with atrial fibrillation.[12] Because we are able to measure the updated CHA$_2$DS$_2$-VASc score risk factors within observational data, we also created and used indicators for those risk factors as secondary analyses.

Predicting rates of thromboembolic events and major bleeding on dabigatran versus warfarin

Figure 1 shows the steps for obtaining predicted rates of thromboembolism and major bleeding from either randomized controlled trials or cross validated models fit to routine care data.

Predictions from randomized controlled trials

Estimated rates from randomized controlled trials of thromboembolic and major bleeding events within low, medium, and high risk categories were obtained from publications of the RE-LY trial comparing dabigatran to warfarin and the Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation (ARISTOTLE) trial comparing apixaban and warfarin (additional details in the web appendix).[4]

The RE-LY trial published rates of stroke or systemic embolism for participants within three categories of CHADS$_2$ (0-1, 2, ≥3) for participants randomized to receive either warfarin or a 150 mg dose of dabigatran.[16] We applied these rates to the population treated as part of routine care by calculating the CHADS$_2$ score in the routine care population, categorizing initiators of

dabigatran or warfarin into the three risk categories, and then assigning each patient the rate reported from the trial according to exposure and risk stratum.

Although annual rates of major bleeding in the RE-LY trial were published for the overall study population, estimates were not provided for risk categories of the HAS-BLED score.[3] Owing to the lack of published evidence on bleeding event rates stratified by baseline risk from the RE-LY trial, we obtained bleeding event rates within three categories of the HAS-BLED score (0-1, 2, ≥3) in patients randomized to the warfarin arm from publications for a trial comparing a different novel oral anticoagulant agent, apixaban, with warfarin (ARISTOTLE).[17]

We applied the annual rates of major bleeding within low, medium, and high risk categories of HAS-BLED from patients receiving warfarin in the ARISTOTLE trial to initiators of warfarin in routine care, using estimates derived from each of the three definitions of major bleeding used in the trial. Major bleeding in the ARISTOTLE trial was defined using three different criteria: ISTH (International Society on Thrombosis and Haemostasis), TIMI (thrombolysis in myocardial infarction), and GUSTO (global use of strategies to open occluded arteries; details in web appendix).[17-19] To obtain the annual risk of major bleeding for patients receiving dabigatran in our routine care cohort, we multiplied the predicted annual rate of major bleeding within each HAS-BLED risk group for warfarin initiators by the hazard ratio from the RE-LY trial comparing dabigatran with warfarin on risk of major bleeding.[17]

Predictions from models developed in routine care patients

We included known thromboembolism and major bleeding risk factors contributing to the CHADS$_2$, CHA$_2$DS$_2$-VASc, or HAS-BLED scores as predictors in Cox proportional hazards models fitted separately in
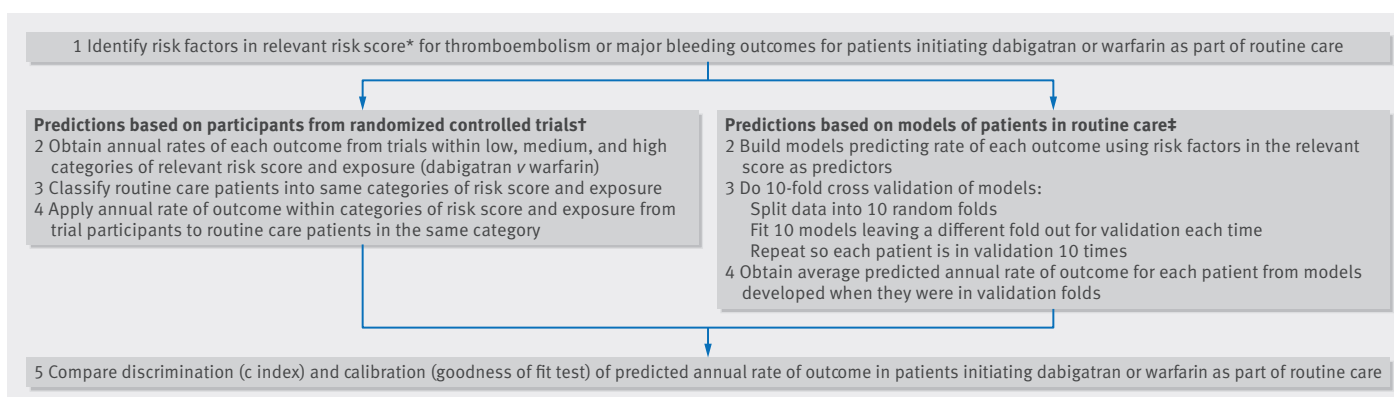

Fig 1 | Steps for obtaining predicted annual rates per 100 patient years of thromboembolism and major bleeding outcomes from randomized controlled trial participants and routine care patients. *Risk score=estimated baseline rate of outcome from relevant external risk score developed in patients not on oral anticoagulation therapy (or mix of treated and untreated). Risk score factors are CHADS$_2$ (congestive heart failure, hypertension, age≥75 years, diabetes mellitus, previous stroke/transient ischaemic attack/thromboembolism (doubled risk weight)), CHA$_2$DS$_2$-VASc (CHADS$_2$ plus vascular disease, age 65-74 years, and female sex (age≥75 years and previous stroke carry doubled risk weight)), and HAS-BLED (age>65 years, hypertension, abnormal renal function, abnormal liver function, prior stroke, bleeding history (or predisposition), drugs predisposing to bleed, alcohol use disorders, and labile international normalized ratio). †Estimated rate of outcomes in trial participants randomized to warfarin or dabigatran. ‡Estimated rate of outcomes from models fit separately in initiators of warfarin and dabigatran in routine care, using risk factors in relevant risk score. Performance evaluated for average estimate from repeated 10-fold cross validation

initiators of dabigatran and warfarin. The models for thromboembolism included sex, age indicators (65-74, age≥75 years), prior stroke/transient ischemic attack/thromboembolism, congestive heart failure, hypertension, diabetes, and vascular disease. The models for major bleeding included indicators for age older than 65 years, hypertension, abnormal renal function, abnormal liver function, prior stroke, bleeding history (or predisposition), drugs predisposing to bleed, alcohol use disorders, and labile international normalized ratio.

Because the performance of models that are evaluated in the data that they are developed can be overly optimistic, we conducted repeated 10-fold cross validation for each model (fig 1) and evaluated performance based on the averaged prediction when patients were in validation folds.[20]

Predictions from baseline risk scores
In addition to use of randomized controlled trials and models developed in routine care patients to predict rates of thromboembolism and major bleeding under treatment with dabigatran or warfarin, we also predicted baseline event rates using published event rates from $CHADS_2$, CHA_2DS_2-VASc, and HAS-BLED risk scores. In the same manner as used for trial predictions, we applied published annual event rates for different levels of each score to the routine care patients with those scores.[1 2 21] For example, if a published event rate for patients with a $CHADS_2$ score of 2 was 4.5 per 100 patient years, then patients in routine care were assigned a predicted event rate of 4.5 per 100 patient years.

These baseline risk scores were originally developed in patients not on oral anticoagulation therapy[1 21] or in a mix of treated and untreated patients.[2] Baseline risk scores do not take into account the effect of therapy on event rates. The purpose of including predictions from baseline risk scores in our comparison was to highlight that event rates from baseline risk scores should be supplemented with risk stratified event rates while on therapy, in order to facilitate evidence based treatment choices (eg, no treatment v treatment A v treatment B).

Performance metrics
We calculated Harrell's c index as a measure of discrimination for survival models. The c index reflects the ability to distinguish patients with the event from those without the event of interest.[22] When a model has discriminative ability no better than chance, the c index will be 0.5, whereas a perfectly discriminating model will have a c index of 1.0.

We assessed the accuracy of our predictions by looking at the calibration of predicted to observed event rates within patients at low, medium, and high baseline risk (eg, 0-1, 2, ≥3) of thromboembolism or major bleeding using categories defined by $CHADS_2$ and HAS-BLED risk scores, respectively. These three categories were used because they were the most granular categories for which we obtained predicted event rates from the randomized controlled trials. We obtained goodness of fit test statistics for calibration of predictions in survival data.[23]

Predicted rates of thromboembolism and major bleeding within low, medium, and high risk categories
We standardized the distribution of baseline characteristics for each exposure group to match that of the entire population with stabilized inverse probability of treatment weights[24] (details in web appendix). We also reported the predicted rate of events for initiators of dabigatran and warfarin using estimates from randomized controlled trials, models developed in routine care patients, and baseline risk scores.

This study was reviewed by the institutional review board at Brigham and Women's Hospital. The cohort was created using SAS version 9.3. Analyses were conducted using Cran R version 3.0.2.

Patient involvement
No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for design or implementation of the study. No patients were asked to advise on interpretation or writing up of results. There are no plans to disseminate the results of the research to study participants or the relevant patient community.

Results
Patient characteristics
In the cohort of 21 934 patients with atrial fibrillation initiating anticoagulation therapy as part of routine care, 30% (n=6516) initiated dabigatran. Annual event rates per 100 patients were 1.7 for thromboembolism and 4.6 for major bleeding. The prevalence of known risk factors for thromboembolism was relatively high (table 1). Over 93% of patients had hypertension and over 25% had type 2 diabetes. Patients initiating warfarin treatment were more likely to be in higher categories of risk for both thromboembolism and major bleeding. These patients were older than dabigatran initiators, more likely to be women, have had prior stroke or transient ischemic attack (20% v 13%), congestive heart failure (27% v 16%), vascular disease (22% v 14%), and more likely to have been admitted to hospital within 30 days before initiation (42% v 32%). Nearly a third (29%) of warfarin initiators had a history of bleeding compared with 16% of dabigatran initiators, and abnormal renal function was more common in warfarin initiators (19% v 8%).

Randomized controlled trial participants were older than patients initiating dabigatran or warfarin as part of routine care, and a similar proportion were female. Trial participants were more likely to have had heart failure and less likely to have diabetes or hypertension than patients initiating warfarin therapy as part of routine care. Both groups had similar proportions of patients with prior stroke or transient ischemic attack. We were not able to compare some of the known characteristics that influence risk of bleeding because these were not reported in the main trial publication.[3] Furthermore, the RE-LY trial excluded patients at higher risk of bleeding.[25] Table 1 presents prevalence of bleeding risk factors from patients randomized to the warfarin arm in the

Table 1 | Characteristics of patients initiating anticoagulation therapy as part of routine care and RE-LY trial participants

Characteristic	Initiators in routine care		RE-LY trial participants*	
	Dabigatran (n=6516)	Warfarin (n=15 418)	Dabigatran (n=6076)	Warfarin (n=6022)
Annual rate of thromboembolic event†	1.2	2.0	1.1	1.7
Annual rate of major bleeding†	2.6	5.5	3.1	3.4
Mean follow-up (months)	5.3 months	4.6 months	About 24 months	About 24 months
Age (years; mean (SD))	62 (11.0)	65 (12.2)	72 (8.8)	72 (8.6)
Age categories				
<65 years	4428 (68.0)	8964 (58.1)	N/A	N/A
65-74 years	1276 (19.6)	3103 (20.1)	N/A	N/A
≥75 years	812 (12.5)	3351 (21.7)	N/A	N/A
Female sex	1776 (27.3)	5195 (33.7)	2236 (36.8)	2213 (36.7)
Congestive heart failure	1019 (15.6)	4122 (26.7)	1934 (31.8)	1922 (31.9)
Type 2 diabetes	1672 (25.7)	4842 (31.4)	1402 (23.1)	1410 (23.4)
Hypertension	6077 (93.3)	14 400 (93.4)	4795 (78.9)	4750 (78.9)
Prior stroke or transient ischemic attack	855 (13.1)	3055 (19.8)	1233 (20.3)	1195 (19.8)
Vascular disease	896 (13.8)	3449 (22.4)	N/A	N/A
Drugs predisposing to bleed‡	1953 (30.0)	4720 (30.6)	2352 (38.7)	2442 (40.6)
Bleeding history (or predisposition)	1066 (16.4)	4480 (29.1)	N/A	N/A (16.7)§
Abnormal liver function	314 (4.8)	988 (6.4)	N/A	N/A (2.7)§
Abnormal renal function	495 (7.6)	2977 (19.3)	N/A	N/A (0.4)§
Alcohol use disorder	163 (2.5)	381 (2.5)	N/A	N/A (2.5)§
Admitted to hospital within 30 days	2090 (32.1)	6494 (42.1)	N/A	N/A
Valvular heart disease	905 (13.9)	3577 (23.2)	0	0
Long term treatment with vitamin K antagonists before study entry	0	0	3049 (50.2)	2929 (48.6)
CHADS$_2$ risk score				
0-1	3419 (52.5)	5825 (37.8)	1958 (32.2)	1859 (30.9)
2	1785 (27.4)	4621 (30.0)	2137 (35.2)	2230 (37.0)
≥3	1312 (20.1)	4972 (32.2)	1981 (32.6)	1933 (32.1)
HAS-BLED risk score				
0-1	2526 (38.8)	4027 (26.1)	N/A	N/A (23.0)§
2	2135 (32.8)	4725 (30.6)	N/A	N/A (36.0)§
≥3	1855 (28.5)	6666 (43.2)	N/A	N/A (41.0)§

Data are no (%) of patients unless otherwise specified.

N/A=information on a baseline characteristic was not available in the publication; SD=standard deviation; CHADS$_2$=congestive heart failure, hypertension, age (≥75 years), diabetes mellitus, and previous stroke/transient ischaemic attack/thromboembolism (doubled risk weight); HAS-BLED=age>65 years, hypertension, abnormal renal and liver function, prior stroke, bleeding history (or predisposition), drugs predisposing to bleed, alcohol use disorders, and labile international normalized ratio.

*RE-LY trial baseline characteristics obtained from trial results (using the 150 mg dose for dabigatran).[3]

†Rate of thromboembolic and major bleeding events for the routine care population reflect estimated rates with no adjustment for confounding. Although these rates are representative of the rates within the patient populations treated with each oral anticoagulation agent, they should not be used to make treatment comparisons without further adjustment, owing to substantial observed differences in baseline characteristics between the patients who initiated dabigatran and patients who initiated warfarin in routine care.

‡For initiators in routine care: NSAIDs, aspirin, clopidogrel, prasugrel, ticagrelor; for RE-LY trial participants: aspirin only.

§Percentages obtained from ARISTOTLE trial participants in warfarin arm.[4]

ARISTOTLE trial,[4] because this is the population from whom we obtained stratified estimated rates of bleeding while on warfarin. Prevalence of major bleeding risk factors in participants randomized to warfarin was lower than observed in routine care.

Roughly half of the RE-LY trial participants were on long term treatment with vitamin K antagonists before study entry and randomization, whereas we restricted the routine care population to patients with no oral anticoagulant use in the previous 365 days. The trial was conducted in patients with non-valvular atrial fibrillation, whereas in routine care, 14% of dabigatran initiators and 23% of warfarin initiators had valvular heart disease.

Discrimination

The c indices for discrimination based on the randomized controlled trial estimates were 0.59 and 0.66 for thromboembolism in dabigatran and warfarin initiators, respectively, and 0.60 for major bleeding outcomes irrespective of the criteria used to define major bleeding (table 2). The discrimination for model based predictions was higher for warfarin initiators than dabigatran initiators (thromboembolism: 0.70 v 0.52 using CHADS$_2$ risk factors, 0.72 v 0.54 using CHA$_2$DS$_2$-VAScrisk factors; major bleeding: 0.64 v 0.58). The lower performance in dabigatran initiators can be attributed in part to the smaller sample size and fewer events available for model development and validation. CHADS$_2$, CHA$_2$DS$_2$-VASc, and HAS-BLED baseline risk scores were not designed to distinguish risk for patients on different treatments, and had discrimination ranging between 0.58 and 0.67 for thromboembolism and 0.60 and 0.62 for major bleeding.

Calibration

Figure 2 depicts the difference between predicted and observed rates per 100 patient years within low, medium, and high risk categories of the CHADS$_2$ score for the thromboembolism outcome and the HAS-BLED

Table 2 | Discrimination and goodness of fit tests for predicted rates per 100 patient years of thromboembolism and major bleeding in initiators of dabigatran or warfarin as part of routine care

| | Dabigatran initiators | | | | Warfarin initiators | | | |
| | | Calibration | | | | Calibration | | |
	Discrimination (c index (95% CI))	Goodness of fit test statistic	P		Discrimination (c index (95% CI))	Goodness of fit test statistic	P	
Thromboembolism*†								
CHADS$_2$ (RCT)	0.66 (0.57 to 0.75)	4.44	0.11		0.59 (0.53 to 0.63)	3.39	0.18	
CHADS$_2$ (model)	0.52 (0.44 to 0.75)	3.99	0.14		0.70 (0.64 to 0.75)	1.40	0.50	
CHADS$_2$ (score)	0.67 (0.57 to 0.77)	244.79	0.00		0.58 (0.52 to 0.63)	584.65	0.00	
CHA$_2$DS$_2$-VASc (model)	0.54 (0.49 to 0.77)	4.23	0.12		0.72 (0.66 to 0.77)	1.62	0.44	
CHA$_2$DS$_2$-VASc (score)	0.66 (0.57 to 0.75)	53.52	0.00		0.63 (0.58 to 0.67)	186.66	0.00	
Major bleeding*†‡								
ISTH (RCT)	0.60 (0.54 to 0.66)	4.83	0.09		0.60 (0.57 to 0.64)	26.89	0.00	
TIMI (RCT)	0.60 (0.54 to 0.66)	5.83	0.05		0.60 (0.54 to 0.66)	56.46	0.00	
GUSTO (RCT)	0.60 (0.54 to 0.66)	5.77	0.06		0.60 (0.54 to 0.66)	67.92	0.00	
HAS-BLED (model)	0.58 (0.50 to 0.68)	3.52	0.17		0.64 (0.61 to 0.68)	6.78	0.03	
HAS-BLED (score)	0.60 (0.54 to 0.67)	6.30	0.04		0.62 (0.59 to 0.66)	36.97	0.00	

RCT=estimated rate of outcomes in trial participants randomized to warfarin or dabigatran; model=estimated rate of outcomes from models fit separately in initiators of warfarin and dabigatran in routine care, using risk factors in relevant risk score; score=estimated baseline rate of outcome from relevant external risk score developed in patients not on oral anticoagulation therapy (or mix of treated and untreated patients).

Performance evaluated for average estimate from repeated 10-fold cross validation; 95% confidence intervals calculated from percentiles of 1000 bootstrap samples.

*No of events of thromboembolism: 33 (dabigatran) and 114 (warfarin); no of events of major bleeding: 75 (dabigatran) and 320 (warfarin).

†Risk score factors: CHADS$_2$=congestive heart failure, hypertension, age (≥75 years), diabetes mellitus, and previous stroke/transient ischaemic attack/thromboembolism (doubled risk weight); CHA$_2$DS$_2$-VASc=factors in CHADS$_2$ plus vascular disease, age 65-74 years, and female sex (age≥75 years and previous stroke carry doubled risk weight); HAS-BLED=age>65 years, hypertension, abnormal renal and liver function, prior stroke, bleeding history (or predisposition), drugs predisposing to bleed, alcohol use disorders, and labile international normalized ratio.

‡Criteria for ascertaining major bleeding in clinical trials: ISTH=International Society on Thrombosis and Haemostasis; TIMI=thrombolysis in myocardial infarction; GUSTO=global use of strategies to open occluded arteries.

score for the major bleeding outcome. Here, differences closer to zero reflect more accurate prediction of event rates within risk categories. Overall, the predicted rates of thromboembolism per 100 patient years from randomized controlled trials and models fit the observed rates in low, medium, and high baseline risk categories similarly (P>0.10; table 2). As would be expected, estimated rates per 100 patient years from baseline risk scores (CHADS$_2$ and CHA$_2$DS$_2$-VASc) predicting the rate of events had the patients not been on oral anticoagulant therapy fit the observed rates poorly, overestimating the rate of thromboembolism in dabigatran and warfarin initiators (P<0.001).

Predictions from randomized controlled trials consistently underestimated the rate of major bleeding observed in warfarin initiators (P<0.001); in the most extreme case, by up to 4.0 major bleeds per 100 patient years in patients with HAS-BLED scores of 3 or higher. The trial predictions also fit poorly in dabigatran initiators using each of the different major bleeding criteria (P values ranging from 0.05 to 0.09). Predicted baseline rates per 100 patient years from the HAS-BLED score also underestimated the observed rate of major bleeding in patients who initiated warfarin (P<0.001). Model based predictions had the best fit for the observed rate of major bleeding per 100 patient years in both dabigatran and warfarin initiators (table 2).

Tables 3 and 4 shows the predicted rates of thromboembolism and major bleeding per 100 patient years after adjustment for confounders. The confounders used for adjustment are listed in the web appendix. Table 3 shows how much higher the anticipated thromboembolism event rates would be if patients remain untreated (baseline risk scores CHADS$_2$ and CHA$_2$DS$_2$-VASc) than if they were to initiate either dabigatran or warfarin. The model based and trial based predictions for rate of thromboembolism per 100 patient years in dabigatran and warfarin treated patients were similar across low, medium, and high risk categories. Although the predicted rates of major bleeding per 100 patient years varied depending on the criteria for defining major bleeding events used in the randomized controlled trial, each criterion predicted similar event rates for dabigatran and warfarin treated patients. By contrast, model based predicted rates of major bleeding were higher in warfarin initiators than in dabigatran initiators; this difference in rate of bleeding was greatest in patients at high baseline risk.

Results for supplementary analyses using a combined thromboembolism outcome that excluded transient ischemic attack, pulmonary embolism, and deep vein thrombosis are available in the web appendix.

Discussion
Principal results
In this study of patients initiating warfarin or dabigatran as part of routine care, use of estimates from the RE-LY randomized controlled trial[3] performed as well as models developed within a routine care population at predicting the rate of thromboembolism per 100 patient years under dabigatran or warfarin treatment. Trial estimates for major bleeding consistently underestimated bleeding risk for patients treated as part of routine care. The poor calibration of the trial estimated rate of bleeding to the observed rate in the routine care population is probably related to several factors, including enrollment of a high proportion of participants with demonstrated tolerance to warfarin and systematic exclusion of patients with conditions that put them at higher risk of bleeding. These factors resulted in distributions of
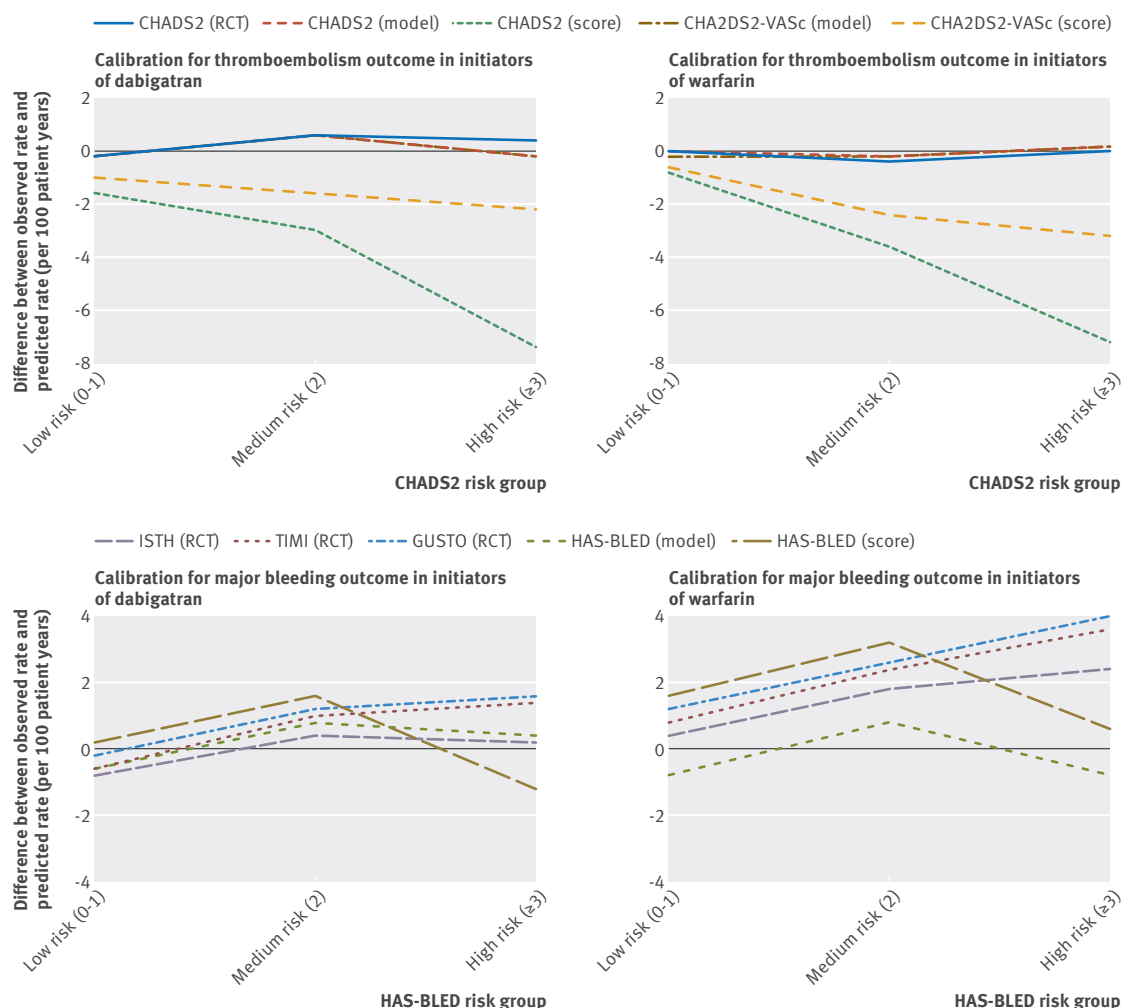
Fig 2 | Calibration of predicted rates per 100 patient years of thromboembolism or major bleeding in initiators of dabigatran or warfarin. Difference between rates calculated as observed rates minus predicted rates. RCT=estimated rate of outcomes in trial participants randomized to warfarin or dabigatran; model=estimated rate of outcomes from models fit separately in initiators of warfarin and dabigatran in routine care, using risk factors in relevant risk score; score=estimated baseline rate of outcome from relevant external risk score developed in patients not on oral anticoagulation therapy (or mix of treated and untreated patients). Performance evaluated for average estimate from repeated 10-fold cross validation. Risk score factors: CHADS$_2$=congestive heart failure, hypertension, age (≥75 years), diabetes mellitus, and previous stroke/transient ischaemic attack/thromboembolism (doubled risk weight); CHA$_2$DS$_2$-VASc=factors in CHADS$_2$ plus vascular disease, age 65-74 years, and female sex (age≥75 years and previous stroke carry doubled risk weight); HAS-BLED=age>65 years, hypertension, abnormal renal and liver function, prior stroke, bleeding history (or predisposition), drugs predisposing to bleed, alcohol use disorders, and labile international normalized ratio. Criteria for ascertaining major bleeding in clinical trials: ISTH=International Society on Thrombosis and Haemostasis; TIMI=thrombolysis in myocardial infarction; GUSTO=global use of strategies to open occluded arteries

Table 3 | Predicted rates of thromboembolism per 100 patient years in patients initiating dabigatran or warfarin as part of routine care

Risk group	RCT		CHADS$_2$ model*†		CHADS$_2$ risk score	CHA$_2$DS$_2$-VASc model*†		CHA$_2$DS$_2$-VASc risk score
	Dabigatran	Warfarin	Dabigatran	Warfarin		Dabigatran	Warfarin	
All	1.1	1.6	1.2 (0.7 to 1.8)	1.2 (1.0 to 1.6)	5.1	1.2 (0.7 to 1.8)	1.2 (1.0 to 1.6)	3.3
Low	0.7	1.1	0.7 (0.4 to 1.1)	0.9 (0.6 to 1.2)	2.0	0.7 (0.4 to 1.1)	0.9 (0.6 to 1.2)	1.6
Medium	0.8	1.4	0.9 (0.5 to 1.4)	0.9 (0.7 to 1.2)	4.5	0.9 (0.5 to 1.3)	0.9 (0.7 to 1.2)	3.3
High	1.9	2.7	2.1 (1.1 to 3.6)	2.1 (1.6 to 2.7)	10.0	2.2 (1.1 to 3.7)	2.1 (1.5 to 2.7)	5.7

RCT=estimated rate of outcomes in trial participants randomized to warfarin or dabigatran; model=estimated rate of outcomes from models fit separately in initiators of warfarin and dabigatran in routine care, using risk factors in relevant risk score; score=estimated baseline rate of outcome from relevant external risk score developed in patients not on oral anticoagulation therapy (or mix of treated and untreated patients).
Performance evaluated for average estimate from repeated 10-fold cross validation.
*95% confidence intervals calculated from percentiles of 1000 bootstrap samples.
†Risk score factors: CHADS$_2$=congestive heart failure, hypertension, age (≥75 years), diabetes mellitus, and previous stroke/transient ischaemic attack/thromboembolism (doubled risk weight); CHA$_2$DS$_2$-VASc=factors in CHADS$_2$ plus vascular disease, age 65-74 years, and female sex (age≥75 years and previous stroke carry doubled risk weight).

Table 4 | Predicted rates of major bleeding per 100 patient years in patients initiating dabigatran or warfarin as part of routine care

Risk group	RCT (ISTH)* D	RCT (ISTH)* W	RCT (TIMI)* D	RCT (TIMI)* W	RCT (GUSTO)* D	RCT (GUSTO)* W	HAS-BLED model†‡ D	HAS-BLED model†‡ W	HAS-BLED risk score‡
All	3.3	3.5	2.6	2.7	2.3	2.5	2.5 (1.9 to 3.3)	4.0 (3.4 to 4.5)	3.5
Low	2.0	2.2	1.7	1.9	1.4	1.5	1.3 (0.9 to 1.9)	2.5 (2.0 to 3.0)	1.0
Medium	3.0	3.2	2.4	2.5	2.2	2.4	2.0 (1.5 to 2.6)	3.1 (2.6 to 3.6)	1.9
High	4.4	4.7	3.3	3.5	3.0	3.3	3.7 (2.4 to 5.4)	5.8 (4.9 to 6.8)	6.6

D=dabigatran, W=warfarin; RCT=estimated rate of outcomes in trial participants randomized to warfarin or dabigatran; model=estimated rate of outcomes from models fit separately in initiators of warfarin and dabigatran in routine care, using risk factors in relevant risk score; score=estimated baseline rate of outcome from relevant external risk score developed in patients not on oral anticoagulation therapy (or mix of treated and untreated patients).
Performance evaluated for average estimate from repeated 10-fold cross validation
*Criteria for ascertaining major bleeding in clinical trials: ISTH=International Society on Thrombosis and Haemostasis; TIMI=thrombolysis in myocardial infarction; GUSTO=global use of strategies to open occluded arteries.
†95% confidence intervals calculated from percentiles of 1000 bootstrap samples.
‡Risk score factor: HAS-BLED=age>65 years, hypertension, abnormal renal and liver function, prior stroke, bleeding history (or predisposition), drugs predisposing to bleed, alcohol use disorders, and labile international normalized ratio.

patient characteristics and range in baseline risks quite different from what might be observed in routine care. That the RE-LY trial included "persistent" users further speaks to the difficulty in trying to use trial data to inform real world practice decisions. In addition, predicted rates of major bleeding within low, medium, and high risk categories were not available from a single head-to-head trial and had to be interpolated using published information from two trials. This interpolation highlights the practical difficulty for healthcare providers using trial results to inform treatment decisions.

Clinical implications

Randomized controlled trials are an essential method for determining the efficacy of medications, and healthcare professionals regularly rely on the results of trials to make important treatment decisions for individual patients. However, trials often are conducted in restricted populations with characteristics that differ from those of patients treated in routine practice. In addition, an average estimate found in a trial might mask considerable heterogeneity in treatment effect.[26] Our results suggest that, once trials have determined efficacy of a drug, analyses of observational data of patients treated in routine practice can be of great value for providing tailored estimates of risk for relevant benefit or safety outcomes under alternative treatment strategies.

Our results are most useful for a clinician and patient dyad trying to choose among no treatment, dabigatran, or warfarin when they know that the patient has certain risk factors included in the established $CHADS_2$ and HAS-BLED risk scores. From these two scores, they have a sense of expected event rates if untreated, but ideally, to inform their decision, they would like to know the expected event rates if the patient were to initiate dabigatran or warfarin. Two potential sources of this information are randomized controlled trials, which is the status quo, and observational data, which are becoming increasingly available. Informed by the

most accurate evidence available, physician-patient dyads can make patient centered shared decisions regarding treatment.

There are some situations in which randomized controlled trials are the only source of evidence to inform decision making. For example, in the case of new drug treatments entering the market, there will be a period during which there are insufficient data from clinical practice to be able to develop and validate prediction models for relevant outcomes, especially if the outcomes are also rare. However, there are many situations in which it is difficult to obtain estimates of risk from trials for important benefit or safety outcomes given clinically relevant treatment alternatives. For instance, when there is no risk score available on which to stratify baseline outcome probability, there are no head-to-head trials comparing relevant treatment alternatives on outcomes of interest, estimates of absolute risk differences are available only for the entire trial population and not within a specific risk score range, and trial inclusion and exclusion criteria result in major differences between trial and clinical practice populations.

Although other studies have previously found that randomized controlled trials underestimate bleeding risk in patients with atrial fibrillation,[27] our study highlights the potential use of secondary, electronically captured patient data to develop risk models to guide treatment decision making. Our results suggested that if physicians and patients relied on predicted bleeding rates based on data from the trial, they might substantially underestimate bleeding risk. In situations where model based predictions can be demonstrated to be more accurate for patients treated as part of routine care, they may prefer to use these predicted event rates as one factor in the decision making process. Using large databases to develop and validate models, which can be iteratively updated over time, can provide stratified estimates of absolute risk to guide decision making and enhance patient care. The performance of such models can be internally cross validated, validated within data source over time, or validated across data sources. Estimates from models based on populations treated in routine care settings are naturally calibrated to a clinical population in which predicted individual risks could be used to guide clinical care. Risk models developed within large observational data sources also have the benefit of being able to be adaptively updated as practice patterns change, indications for use expand, or new treatments emerge for which there are no head-to-head trials.

Study limitations

Our study was limited by the focus on dabigatran and warfarin, without evaluation of the other novel oral anticoagulation agents that have recently entered the market. The relatively small number of patients in our observational data who initiated dabigatran in routine care and had thromboembolic outcomes during the study limited the reliability of the model based predictions for patients exposed to dabigatran. However,

with larger data sources, and accrual of greater experience with dabigatran, one would expect that model based predictions would become increasingly more reliable while the trial based predictions would not change.

Our approach was also limited by the inability to capture important clinical variables that are not typically measured in electronic claims data (eg, smoking, body mass index). Under-recording of HAS-BLED risk factors in observational data could result in higher observed rates of bleeding within HAS-BLED risk categories owing to patients who are misclassified into lower risk categories. However, the greatest differences between the randomized controlled trial data and observational data predictions were in the high risk HAS-BLED category (score≥3). These patients would not have been assigned a higher risk category had more HAS-BLED risk factors been captured in observational data. Furthermore, in spite of potential under-recording of risk factors in observational data, it is likely that that the average HAS-BLED score for routine care patients at high baseline risk of bleeding was still higher than the average score in trials participants with a HAS-BLED score of 3 or more. However, the mean score within HAS-BLED risk categories was not reported from the trials, so this cannot be evaluated.

The median duration of follow-up while on treatment was shorter in routine care (median<6 months) than in the randomized controlled trials (median 24 months), and half of trial participants were not naive to oral anticoagulation at the time of randomization. The higher rates of major bleeding observed in routine care patients could be partly attributed to exclusive contribution of person time from the first months following initiation of oral anticoagulation therapy. By contrast, rates from the trials included person time from participants with demonstrated tolerance for anticoagulants as well as person time further from initiation of treatment than was observed in the routine care patients.

Another limitation was how outcomes were defined in the randomized controlled trials versus the observational data. Trial outcomes were based on prespecified and adjudicated criteria whereas outcomes in the observational data were defined on the basis of previously validated claims algorithms. These differences in outcome assessment could have contributed to the differences in rates. Furthermore, there were concerns raised by the US Food and Drug Administration and others regarding how major bleeding events were ascertained and counted in the RE-LY trial.[28] Although targeted reviews did not result in changes to the distribution of major bleeding across study arms that altered the main findings, misclassified events could contribute to the poor calibration of trial based predictions to bleeding in routine care patients.

There was also potential for residual confounding in the reported adjusted rates of thromboembolism or major bleeding for dabigatran and warfarin within the risk categories of the CHADS$_2$ and HAS-BLED scores. The stabilized inverse probability weights we used only adjust for the covariates that were included in the estimation of the weights, and there could be residual confounding from confounders that were not included.

Finally, in spite of our repeated cross validation, it is possible that the model based estimates were still over fit to the data, making performance of the model based estimates overly optimistic. The performance of model based predictions could be quite different if applied to a different large healthcare database that covers a different patient mix. Although model based predictions might not be as useful to patients outside of the type of patients found in a specific large healthcare data source, they can still be quite accurate and of considerable benefit for patients who are captured within that large healthcare data source. Furthermore, the process we have described can be readily applied to regularly update estimates over time and obtain tailored estimates within other large healthcare data sources.

Conclusion

Currently, risk scores are used to guide care for patients with atrial fibrillation, aiding with the decision of whether to initiate oral anticoagulation therapy, but not which oral anticoagulant to select. Risk stratified event rates while on alternative treatments can be used to enhance informed decision making. Although no risk score is perfect, more accurate estimates of event rates can help prescribers and patients make better informed decisions and improve patient centered care. In our example, models developed in routine care patients provided accurate, tailored estimates of risk and benefit under alternative oral anticoagulation therapies.

terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/.