

Can Machine Learning from Real-World Data Support Drug Treatment Decisions? A Prediction Modeling Case for Direct Oral Anticoagulants

Medical Decision Making
2022, Vol. 42(5) 587–598
© The Author(s) 2021



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0272989X211064604
journals.sagepub.com/home/mdm



Andreas D. Meid , Lucas Wirbka, ARMIN Study Group,
Andreas Groll, and Walter E. Haefeli 

Background: Decision making for the “best” treatment is particularly challenging in situations in which individual patient response to drugs can largely differ from average treatment effects. By estimating individual treatment effects (ITEs), we aimed to demonstrate how strokes, major bleeding events, and a composite of both could be reduced by model-assisted recommendations for a particular direct oral anticoagulant (DOAC). **Methods:** In German claims data for the calendar years 2014–2018, we selected 29 901 new users of the DOACs rivaroxaban and apixaban. Random forests considered binary events within 1 y to estimate ITEs under each DOAC according to the X-learner algorithm with 29 potential effect modifiers; treatment recommendations were based on these estimated ITEs. Model performance was evaluated by the c-for-benefit statistics, absolute risk reduction (ARR), and absolute risk difference (ARD) by trial emulation. **Results:** A significant proportion of patients would be recommended a different treatment option than they actually received. The stroke model significantly discriminated patients for higher benefit and thus indicated improved decisions by reduced outcomes (c-for-benefit: 0.56; 95% confidence interval [0.52; 0.60]). In the group with apixaban recommendation, the model also improved the composite endpoint (ARR: 1.69 % [0.39; 2.97]). In trial emulations, model-assisted recommendations significantly reduced the composite event rate (ARD: –0.78 % [–1.40; –0.03]). **Conclusions:** If prescribers are undecided about the potential benefits of different treatment options, ITEs can support decision making, especially if evidence is inconclusive, risk-benefit profiles of therapeutic alternatives differ significantly, and the patients’ complexity deviates from “typical” study populations. In the exemplary case for DOACs and potentially in other situations, the significant impact could also become practically relevant if recommendations were available in an automated way as part of decision making.

Highlights

- It was possible to calculate individual treatment effects (ITEs) from routine claims data for rivaroxaban and apixaban, and the characteristics between the groups with recommendation for one or the other option differed significantly.
- ITEs resulted in recommendations that were significantly superior to usual (observed) treatment allocations in terms of absolute risk reduction, both separately for stroke and in the composite endpoint of stroke and major bleeding.
- When similar patients from routine data were selected (precision cohorts) for patients with a strong recommendation for one option or the other, those similar patients under the respective recommendation showed a significantly better prognosis compared with the alternative option.
- Many steps may still be needed on the way to clinical practice, but the principle of decision support developed from routine data may point the way toward future decision-making processes.

Corresponding Author

Andreas D. Meid, Department of Clinical Pharmacology and Pharmacoeconomics, Heidelberg University Hospital, Im Neuenheimer Feld 410, Heidelberg, 69120, Germany; (andreas.meid@med.uni-heidelberg.de).

Keywords

claims data, clinical decision support system, direct oral anticoagulants, heterogeneous treatment effects, machine-learning, personalized medicine

Date received: June 15, 2021; accepted: November 12, 2021

Background

When choosing between several drug treatments for their patients, physicians and health care professionals are often unsure which option is “best.”^{1,2} Generalization of guidelines or evidence from randomized controlled trials (RCTs) cannot overcome this uncertainty because they validly estimate average treatment effects between groups in a study population but cannot explain potentially heterogeneous treatment effects of single individuals.^{3,4} Thus, patient characteristics can modulate average effects and lead to different individual treatment effects (ITEs) at the patient level.⁵ Good assessment of ITEs is critical when physicians seek to provide individualized care to their patients,^{2,6} especially in situations in which significant benefits must be weighed against serious harms of treatment options, such as the prevention of thromboembolic events in patients with nonvalvular atrial fibrillation (AF) with direct oral anticoagulants (DOACs) and the associated bleeding risk.^{7,8} Here, treatment decisions could become more precise if real-world evidence was processed using machine learning to provide personalized recommendations for decision making with DOACs. If the incidence of stroke and (major) bleeding events could be reduced through model-assisted decision making, both the individual patients who are offered the potentially most favorable option and also the health care systems as a whole would benefit significantly.

The current literature shows promising examples from RCT reanalyses to infer ITEs by machine learning⁹ or from observational studies to elucidate effect modifiers that modulated the (individual) effectiveness of different classes of antidiabetics.¹⁰ ITEs within a pharmacological class such as DOACs have not been investigated yet. Until recently, it was generally assumed in clinical practice that DOACs were largely equivalent. This represents a favorable situation in which the potential advantages of a model-assisted recommendation are due to the fact that the decision algorithm is free of existing guidelines and reasonably uninfluenced by firm beliefs of prescribing physicians. Nevertheless, at least the averaged means of the options rivaroxaban and apixaban appeared to be superior to each other under different circumstances: slight advantages were partly apparent for stroke prevention under apixaban^{11–14} or rivaroxaban^{7,8,15–18} supposedly at the expense of a higher bleeding risk with rivaroxaban.^{7,8,11–13,17,18} This emphasizes how important a benefit–risk assessment is for each individual patient. Thus, it appears obvious to explore ITEs in terms of effectiveness and harm, which was generally very rarely found in a systematic review evaluating the personalization of benefit and harm results from RCTs, in which only 1 analysis included a clinical prediction guide.¹⁹

The situation in which a treatment effect is modulated by a single or a few patient characteristics can be well approached using conventional regression methods. In more complex situations with often high-dimensional individual characteristics, this is no longer readily possible, for example, when many modulators are present that also influence each other (possibly even in a nonlinear manner).²⁰ With the availability of Big Data from large claims databases and electronic health records, machine learning is able to efficiently train predictive models based on input features (predictors) and observed outcomes to predict whether a certain outcome would occur for a patient.^{21,22} However, ITEs are the difference between the outcomes of 2 (or more) treatments. In real-world health care data, these outcomes typically cannot be observed across different treatment options for the same patient and are therefore not available as a direct intraindividual comparison suitable for training prediction models. Using a causal inference framework, one can nevertheless predict outcomes in patients as if they had received one

Department of Clinical Pharmacology and Pharmacoepidemiology, University of Heidelberg, Heidelberg, Germany (ADM, LW, WEH). Department of Statistics, TU Dortmund University, Dortmund, Germany (AG). A list of ARMIN investigators is provided in the Acknowledgements section at the end of this article. The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support for this study was provided in part by a contract with the statutory health insurance fund *AOK PLUS* (funding of Lucas Wirbka). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. Andreas D. Meid is funded by the Physician-Scientist Programme of the Medical Faculty of Heidelberg University. The funding body did not play any role in the design of the study or in the data collection, analysis and interpretation of data, or writing of the manuscript.

option or the other option and then calculate their difference.^{2,21} Following this split-model approach, random forest-based methods can be used, and their predictive accuracy must be evaluated out of sample, taking into account their potential for overfitting.

In this proof-of-concept study, we used data from a large health insurance company and explored the potential for machine-learning techniques to infer the individual risk of stroke and major bleeding with rivaroxaban or apixaban and thus also to facilitate individualized recommendations for DOAC therapy. Therefore, we developed and internally validated prediction models for ITEs in an observational cohort of new users of rivaroxaban and apixaban by exploiting clinical characteristics and health system parameters from a processed claims data set. We aimed to answer 3 key questions: 1) Can routinely available information act as meaningful effect modifiers to modulate responses to DOACs? 2) How much could the resulting recommendations have improved outcomes compared with the actual treatment options chosen in the retrospective database? 3) How would such individualized recommendations translate into avoided clinical events in a future implementation of a model-assisted decision-support tool?

Methods

Data Source and Study Population

This study was a predictive modeling study in a previously defined cohort of DOAC patients (referred to henceforth as the “prestudy”) that followed the guidelines for “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis” (TRIPOD; see Supplementary Table S1 in the Supplementary Appendix).²³ The data source completely corresponded to the prestudy, a preparatory analysis of the ARMIN (“Arzneimittelinitiative Sachsen-Thüringen”) program aiming to improve medication quality and safety. In particular, this was an observational study investigating adherence to the DOACs rivaroxaban and apixaban in claims from a large German statutory health insurance company (AOK PLUS) in the calendar years 2014–2018. Our analysis thus built on already processed data (described in detail elsewhere).²⁴ In brief, DOAC-naïve patients with AF who newly initiated either rivaroxaban or apixaban treatment were followed up from their first DOAC prescription if they met the inclusion criterion of at least 3 follow-up prescriptions and had no exclusion criterion in their medical history. Supplementary Figure S1 depicts the flow of participants and how the study

cohort was defined by inclusion and exclusion criteria. The follow-up was in accordance with the intention-to-treat principle, while patients were censored for premature death or DOAC discontinuation with subsequent switch to vitamin K antagonists. All data were fully anonymized for the analysts; in Germany, claims data analyses do not require ethics committee approval by law.

Outcomes

To evaluate the effectiveness of the DOACs, we considered a previously validated and established code set (*International Classification of Diseases*, 10th revision [ICD-10]) to define the outcome stroke in inpatient codes for hospital admission and main diagnoses as the day of the respective hospital admission (i.e., we aimed for higher coding validity than observed for outpatient diagnoses at the expense of potentially missing fatal cases that were not hospitalized; see also Supplementary Table S2 for outcome code sets). Likewise, established ICD-10 codes were applied to detect major bleeding in hospital admission codes as the safety outcome (Supplementary Table S2). A composite endpoint of both stroke or major bleeding was also considered when at least 1 of the separate endpoints occurred within the observation period for model development (binary outcome definition); for the prognosis of treatment recommendations, the first type of event from the separate endpoint was considered (time-to-event outcome definition).

Predictors

At study baseline, when the treatment decision is to be made, a list of 29 variables was available from the prestudy as potential confounders with possible impact on DOAC treatment decisions (Supplementary Table S3). To address the question of whether and which potential confounders modulate DOAC responses in a meaningful way, we considered basic demographics (age, sex), health care services (enrolment into the ARMIN program, prior AF diagnosis in the hospital), comorbidities (Charlson²⁵ and Elixhauser scores²⁶ with distinct comorbidity groups indicating diabetes, hypertension, heart failure, depression, tumors, anemia, or dementia), medical history (occurrence of billing codes in the previous year) of stroke, ischemic heart disease, dyslipidemia, thromboembolism or major bleeding, a CHA₂DS₂-VASc operationalization to indicate risk of thromboembolic events,²⁷ and medication use in the previous year (number of drugs, antihypertensives, antiplatelets, vitamin K antagonists, and lipid-lowering drugs).

Sample Size

Selection of new initiators resulted in 29,901 patients, of whom 16,073 patients were treated with apixaban and 13,828 patients with rivaroxaban. Patients were followed up until their first event or censoring due to death, switching to vitamin K antagonists, or the end of the observation period. All patients were randomly allocated (3:2) to a training set and test to yield samples of 17,722 observations for model development and 12,179 patients for model validation. No missing values existed, because only complete claims were available.

Statistical Analysis

Model development: Estimation of ITEs for treatment recommendations. Using the causal inference framework,² we estimated the potential outcomes of each patient under the potential treatment with apixaban and rivaroxaban. The ITE is thus the difference between the individual outcome probabilities under apixaban and rivaroxaban. We referred to this difference as the benefit score in accordance with a recently introduced framework^{6,28} so that positive benefit scores above the decision threshold of 0 indicated recommendations for apixaban. To actually estimate benefit scores, we used the machine-learning algorithm called X-learner.²⁹ Therefore, we considered outcomes within 365 d of follow-up as binary events. Accounting for the fact that patients could have been censored before having the chance to experience an event, we weighted observations according to their inverse probability of being censored³⁰ within the machine-learning techniques of random forests. We used standard random forests to estimate benefit scores for the outcome of major bleeding, while estimation for stroke relied on a tuned random forest³¹ with automated hyperparameter tuning in randomly down-sampled data to increase the proportion of events to nonevents to a 3:7 ratio.³² While this procedure was considered most suitable, sensitivity analyses included also standard (i.e., nontuned) random forests with raw and randomly down-sampled data. To account for potential imbalances in actually received treatments, propensity score weighting was applied to outcome probabilities under each treatment option when calculating the benefits scores from their difference. Supplementary Figure S2 visually summarizes the distinct steps of benefit score estimation for the separate outcomes stroke and major bleeding.

For the composite endpoint, we built a generalized linear model predicting the probabilities for the composite endpoint by using the distinct benefit scores for stroke and major bleeding as 2 independent variables.³³

A decision threshold for these probabilities is needed to decide whether a recommendation for either apixaban or rivaroxaban is to be made. We determined this decision threshold in the training data by dividing probabilities into deciles and chose the threshold maximizing the emulated benefit as if the model were implemented into regular care (see the “Model Evaluation” section).

Model evaluation: Validation of personalized recommendations. According to the split-sample approach, we calculated 3 metrics for out-of-sample performance in the test data. In general, it is to evaluate whether our model-assisted recommendations are better than treatment assignment by simple chance while also accounting for the observed treatment allocation in the retrospective data source. First, the c-for-benefit metric considers these triplets of actually received treatment, recommended treatment, and clinical outcome by comparing outcomes in pairs of patients matched on benefit scores but discordant for observed treatment allocation.³⁴ A value greater than 0.5 thus indicates a higher probability than chance that a pair with greater observed benefit in outcomes also has the higher predicted benefit if 2 matched pairs with unequal benefit scores are randomly chosen. Second, absolute risk reductions (ARRs) for the treatment comparison between apixaban and rivaroxaban can be derived from logistic regression models fitted to buckets of test patients with recommendation for apixaban or rivaroxaban.⁹ Thus, a positive ARR is expected to result for apixaban versus rivaroxaban in the bucket of apixaban recommendation, whereas a negative ARR is expected to result for apixaban versus rivaroxaban in the bucket of rivaroxaban recommendation. Third, it is of interest how a model-assisted decision rule could affect outcomes upon implementation into clinical practice and how much outcome frequencies could be reduced in comparison with the observed treatment allocation. While an actual trial would be required to assess the impact (clinical utility), observational data can be used to project the clinical utility by trial emulation.³⁵ This yields absolute risk differences (ARDs; to be compared with 0 as the null value). For all 3 metrics, statistical inference was based on 95% confidence intervals derived by nonparametric bootstrap (250 samples⁹), and significance was defined by noninclusion of the null values in the confidence intervals.

Personalized prognosis: Projected benefit in precision cohorts. One way to visualize and communicate model-assisted treatment recommendations is to present a personalized prognosis of recommended and alternative

treatment options in patients with similar characteristics (i.e., precision cohorts³⁶). Thus, we chose 2 exemplary patients from the whole data source, defined a set of variables to select a precision cohort of similar patients (age, Elixhauser score, Charlson score, diabetes and hypertension with complications, prior stroke, prior major bleeding, CHA₂DS₂-VASc score, ischemic heart disease, dyslipidemia, prior in-hospital AF diagnosis), and calculated the Gower distance metric to select the 25% most similar patients as the precision cohort of each patient.³⁶ Time-to-first-event analysis of the composite endpoint was visualized in Kaplan–Meier plots with statistical inference based on the log-rank test.

Software. Statistical analyses were conducted with the R software environment in version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria) using the key packages *personalized*,²⁸ *ranger*,³⁷ *tuneranger*,³¹ and *SimBaCo*.³⁶

Role of the Funding Sources

The funding bodies did not play any role in the design of the study and data collection, analysis and interpretation of data, or writing of the manuscript.

Results

The groups treated with apixaban and rivaroxaban were largely similar in their patient characteristics (see Table 1). Most importantly, the propensity scores for treatment allocation obtained from these variables showed substantial overlap, which is required for weighting ITEs to calculate benefit scores (Supplementary Figure S3). Generally, all variables were equally distributed by random allocation to training and test data, which also applied to study outcomes occurring at an incidence of 2.0% for stroke and 9.7% for major bleeding (Supplementary Table S4). In the training data used for model development, apixaban users experienced fewer bleeding events (8.9% v. 10.7% in rivaroxaban users) but conversely more strokes (2.2% v. 1.7% in rivaroxaban users).

The developed machine-learning models yielded benefit scores for each patient that were used to make treatment recommendations for each DOAC. These model-assisted recommendations resulted in a substantial proportion of patients being recommended a different option than they actually received. According to the efficacy (stroke) model, 47.6% of rivaroxaban users were recommended to use apixaban, and 63.9% of apixaban users were recommended to use rivaroxaban.

According to the safety (bleeding) model, significantly fewer apixaban users were recommended rivaroxaban (25.0%). Supplementary Figures S4 and S5 visualize treatment recommendations of the efficacy and safety models as a function of study variables, observed outcomes, and treatments received. Groups stratified by their recommendation also differed significantly in most variables when considering the composite endpoint (Supplementary Table S7). Supplementary Table S8 uses the characteristics of 4 patients as examples to show how model-assisted recommendations can result.

When model performance was evaluated in the independent test data set, the stroke model resulted in significantly better decisions with reduced strokes (c-for-benefit: 0.56; 95% confidence interval [0.52; 0.60], see Figure 1A). The safety model for major bleeding less clearly recognized individual differences in bleeding risk (c-for-benefit: 0.52 [0.49; 0.54]). Combining both outcomes to the composite endpoint yielded favorable (raw) ARRs in the bucket of apixaban (ARR: 1.69% [0.39; 2.97]) and rivaroxaban recommendations (ARR: -0.88% [-2.93; 1.21]; Figure 1B). When accounting for outcome frequencies, prescription prevalence, and covariates in an emulated trial, the model-assisted recommendations could significantly have reduced the frequency of (the composite) clinical outcomes (ARD: -0.78% [-1.40; -0.03]; Figure 1C), so that about 1 of 100 events could be additionally avoided by model-assisted treatment recommendations. To visualize the impact for individual patients, we generated precision cohorts with the 25% most similar patients to 2 exemplary patients, 1 of whom received a recommendation for apixaban and 1 a recommendation for rivaroxaban. Time-to-first-event analysis regarding the composite endpoint revealed significant improvements for each recommended treatment over the alternative in the respective precision cohort (Figure 2).

Discussion

Our findings reinforced the assumption that real-world data contain valuable information that can be used to support medical decision making for new patients. Our study is also noteworthy because we compared very closely related pharmacologic alternatives within a class of compounds that are generally considered to be largely equivalent (only the bleeding risk was seen in a more differentiated way³⁸). Following our particular case, personalized recommendations based on ITEs could be obviously suitable to improve current practice by

Table 1. Patient Characteristics in Training Data for Model Development Stratified for DOAC Treatment with Apixaban and Rivaroxaban

	Apixaban (n = 9 547)	Rivaroxaban (n = 8 175)	Total (n = 17 722)
Demographics			
Sex (female), n (%)	5,625 (58.9)	4,529 (55.4)	16,979 (56.8)
Age (mean \pm SD)	79.8 \pm 8.8	77.8 \pm 9.2	78.9 \pm 9.1
Comorbidities and clinical characteristics			
Elixhauser score (mean \pm SD)	7.68 \pm 3.02	7.00 \pm 2.89	7.36 \pm 2.97
Elixhauser groups, n (%)			
Diabetes (uncomplicated)	4,781 (50.1)	3,959 (48.4)	14,698 (49.2)
Diabetes (complicated)	3,281 (34.4)	2,490 (30.5)	9,714 (32.5)
Heart failure	6,346 (66.5)	4,889 (59.8)	18,829 (63.0)
Depression	2,163 (22.7)	1,664 (20.4)	6,535 (21.9)
Hypertension (uncomplicated)	8,920 (93.4)	7,623 (93.2)	28,006 (93.7)
Hypertension (complicated)	4,439 (46.5)	3,296 (40.3)	12,961 (43.3)
Solid tumor	1,326 (13.9)	1,125 (13.8)	4,226 (14.1)
Tumor metastases	204 (2.1)	221 (2.7)	717 (2.4)
Anemia (deficiency)	988 (10.3)	633 (7.7)	2,728 (9.1)
Anemia (blood loss)	158 (1.7)	106 (1.3)	440 (1.5)
Renal disease	4,745 (49.7)	3,221 (39.4)	13,395 (44.8)
Dementia	1,935 (20.3)	1,404 (17.2)	5,613 (18.8)
Charlson score (mean \pm SD)	3.84 \pm 2.12	3.37 \pm 2.07	3.61 \pm 2.11
Ischemic heart disease, n (%)	5,089 (53.3)	4,172 (51.0)	15,631 (52.3)
Dyslipidemia, n (%)	5,690 (59.6)	4,435 (54.3)	17,002 (56.9)
CHA ₂ DS ₂ -VASc score, median (IQR) ^a	5 (4; 5)	4 (3; 5)	4 (4; 5)
Medical history (past 12 mo), n (%)			
Prior stroke	1,710 (17.9)	994 (12.2)	4,559 (15.2)
Prior thromboembolism	239 (2.5)	191 (2.3)	740 (2.5)
Prior bleeding event	2,939 (30.8)	2,306 (28.2)	8,907 (29.8)
Prior hospitalization for atrial fibrillation	7,196 (75.4)	5,821 (71.2)	21,885 (73.2)
Medication (in past 12 mo)			
Antihypertensives, n (%)	8,922 (93.5)	7,555 (92.4)	27,848 (93.1)
Lipid-lowering drugs, n (%)	4,027 (42.2)	3,228 (39.5)	12,166 (40.7)
Antiplatelets, n (%)	2,101 (22.0)	1,619 (19.8)	6,219 (20.8)
Number of different drugs, median (IQR)	10 (7; 14)	9 (6; 13)	10 (6; 14)
Prior vitamin K antagonist treatment, n (%)	3,438 (36.0)	2,896 (35.4)	10,766 (36.0)
Health care utilization, n (%)			
ARMIN program enrolment	215 (2.3)	158 (1.9)	607 (2.0)

IQR, interquartile range.

^aCHA₂DS₂-VASc risk score according to Lip et al.²⁷

reducing the separate events of stroke, major bleeding, or their composite. Our patient variables acting as potential effect modifiers had the strongest impact on the discriminatory ability for stroke benefit, and all are readily available from routine data sources, underlining the practicability of this approach. Model-assisted recommendations were estimated to be superior to actually received treatments, and individualized recommendations could additionally reduce the absolute risk by about 1%, which is substantial within a class of highly effective treatment options,^{38,39} a high prevalence of AF in the population of older people,^{40,41} and a large burden of disease with high annual costs in the year after a stroke.^{42,43}

The specific results of this proof-of-concept study need to be viewed in a broader context. Projects beyond our prototype can use the wealth of longitudinally collected data from statutory health insurers to generate real-world evidence for complex cases without clear guideline recommendations (e.g., multimorbidity, polypharmacy, elderly patients). At all levels of complexity and with treatments affecting endpoints in a differentiated way, adequate support is needed to enable informed decision making for each particular situation and different points of care. Moreover, many decisions do not require a population-based but rather an individualized weighing of benefits and risks. This presents a multidimensional problem for which RCTs do not provide sufficient

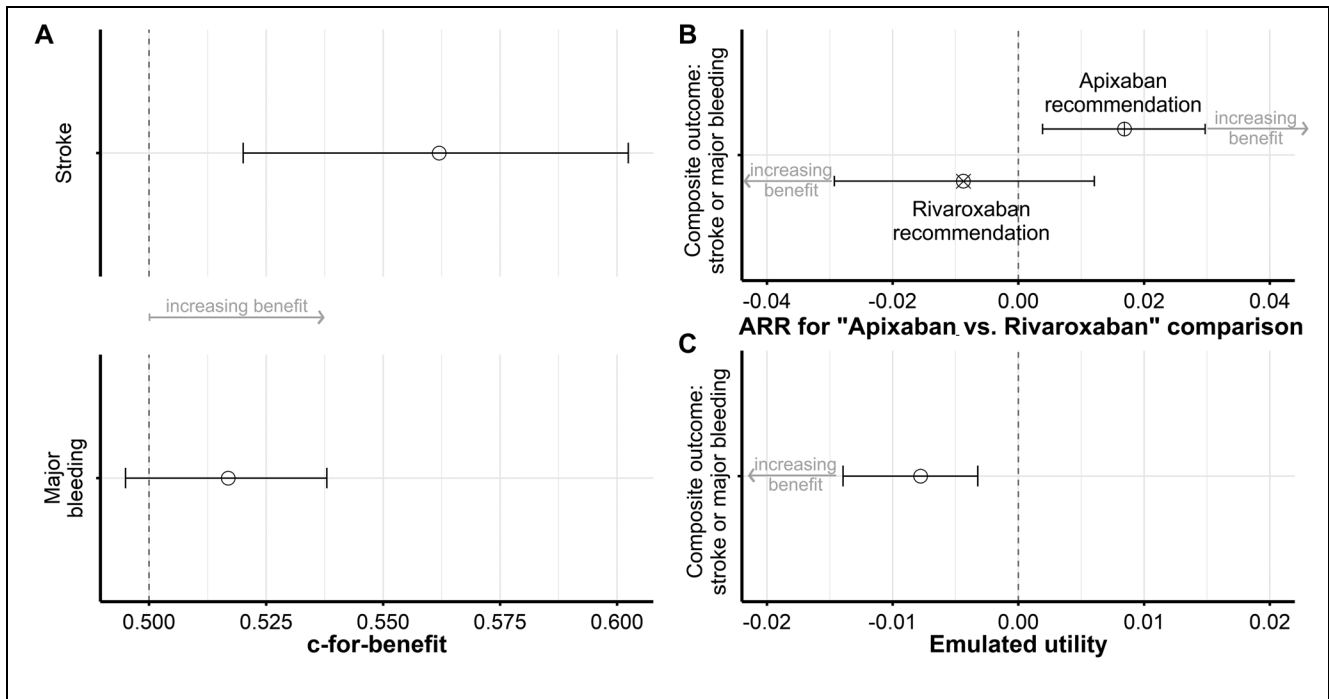


Figure 1 Performance metrics to evaluate the model-assisted treatment recommendations in the test data. (A) The c-for-benefit statistic quantifies the discrimination for benefit considering actually received treatment, recommended treatment, and clinical outcome. (B) Absolute risk reductions (ARRs) refer to the group comparison “apixaban v. rivaroxaban” in buckets of patients with a recommendation for apixaban (⊗) or rivaroxaban (⊗). (C) Emulated utility is expressed as the absolute risk difference over standard of care upon a potential implementation of the model-assisted recommendation into a decision-support system in clinical practice.

information.³ A necessarily personalized decision, on the other hand, requires a multilayered information base that contain all constellations of patient characteristics as they occur in everyday medical practice.⁹ Ultimately, decision making can be based on models developed in such informative data sources and thus enable personalized medicine. Through model-assisted recommendations, the prescribing physician (and patient in a shared decision-making process) would not have to rely solely on symptoms, generalizing guidelines, or personal experience. Instead, he or she could draw from the wealth of such everyday experiences from routine data while accounting for interindividual variability in individual patient characteristics.

Possible explanations can go in many directions. Considering the time frame of the database from 2014–2018, many direct comparisons in real-world data were not yet available,^{7,8,12,13} so that little guiding evidence beyond pivotal trials³⁹ was available in the broad field of outpatient care. Thus, the 2 drugs were considered to be practically equivalent, with no pronounced prescription preferences in general (apart from the individual prescriber). This was evident, for example, because there were no

striking differences between observed patient allocation to apixaban or rivaroxaban (Table 1). Such a situation was certainly advantageous for predicting ITEs to derive personalized recommendations. Indeed, the differences in group means of variables were more pronounced when groups were derived from model-assisted recommendations (see Supplementary Table S7). These differences reflected the risk-benefit tradeoff for the question who can have an option recommended with acceptable risk of bleeding and good chance of response. We would like to emphasize that the single-variable differences should be regarded as descriptive at best, because the ultimate probability of treatment success (benefit score) depends on the interaction of all variables. In this regard, the stroke models performed better than the bleeding models (see Figure 1A), possibly because many co-medications (including temporary or over-the-counter drugs) were not available in the present variables.⁴⁴ Nevertheless, the combined consideration of the composite endpoint (stroke and major bleeding) was impressively successful after weighting both aspects (and the sheer number of far more bleeding events than strokes may certainly have contributed to the significant result).

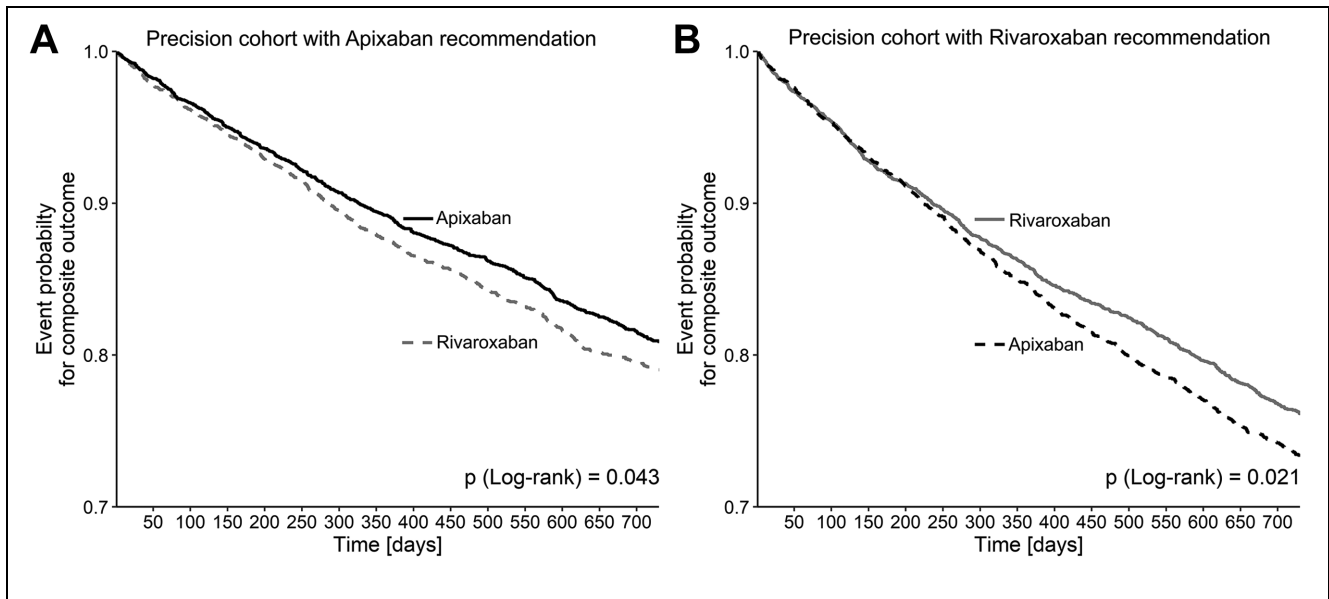


Figure 2 Kaplan-Meier plots in precision cohorts in a sample with similar characteristics to a patient being recommended apixaban (A, left) or a sample with similar characteristics to a patient being recommended rivaroxaban (B, right). The patient on apixaban recommendation is 79 years old, is assigned an Elixhauser comorbidity score of 7 and a Charlson score of 3, has not been diagnosed with complications for diabetes or hypertension, has already experienced both a stroke and a major bleeding event, has a CHA₂DS₂-VASc score of 4 with an in-hospital diagnosis for atrial fibrillation, and has a diagnosis for ischemic heart disease but no dyslipidemia or renal disease. The patient on rivaroxaban is 84 years old with an Elixhauser comorbidity score of 16 and Charlson score of 8, has experienced complications for diabetes and hypertension, had no prior stroke but a major bleeding event in his medical history, has a CHA₂DS₂-VASc score of 5 with an in-hospital diagnosis for atrial fibrillation, and is diagnosed for ischemic heart disease, dyslipidemia, and renal disease. The 25% most similar patients in each case are followed up according to the treatments received (black lines: apixaban; gray lines: rivaroxaban), with solid lines indicating the respective recommendation for the patient from which the precision cohort was formed.

Although our innovative approach overlaps only in small parts with previous work, consensus criteria for treatment-guiding risk modeling can be discussed.⁴⁵ Although DOACs are generally considered to be highly efficacious (with subtle perceived differences), they still have a nonnegligible risk of side effects. Such a situation warrants exploration of individual probabilities beyond generalized population-level probabilities of treatment benefit.² This appears especially promising with regard to the heterogeneous group of AF patients who are generally older and present with a large case-mix variability.⁴¹ Although there are many examples of prognostic models for stroke²⁷ or bleeding⁴⁴ in AF patients, it is not clear how these models can be used for (within-class) DOAC treatment decision making (let alone combined consideration of both aspects). For this, our approach provides a straightforward path using well-established variables that have previously been shown to be prognostic or determine treatment decisions⁴⁶ and are largely available in routine clinical care.

Limitations relate primarily to the data source with available sample sizes. The limitations of routine data are well known, as they are primarily created for billing purposes and cannot be directly used for research purposes. Therefore, established and validated code sets were used throughout the study to define potential effect modifiers and outcomes, and these definitions were also made transparent (see Supplementary Appendix) to allow replications of the work. Linked to the basis of the data are also the sample size and the number of events. It is known that machine-learning methods require large samples for development⁴⁷; however, for validation data sets, it is equally important to have a sufficient number of events (of note, our data set clearly met the prerequisite of at least 100 events).⁴⁸ For this proof-of-concept study, the easy-to-understand split-sample approach worked satisfactorily, as evidenced by the good performance of the down-sampled data set. Nevertheless, alternative methods (e.g., cross-validation or bootstrapping) could be used for other applications of our principle,

especially with small samples.⁴⁸ Furthermore, the modified intention-to-treat approach of the given data basis should be noted, where a certain persistence of at least 3 prescriptions was also assumed. This is a special feature from the prestudy, in which treatment switches were also considered as censoring. Restricting the analysis to 2 exclusive substances administered to continuously insured patients may also have introduced selection bias. This necessary feature of the study design does not represent the full complexity of health care reality, from which less standardized cases could not be used. Conversely, it minimizes the potential impact of differences in insurance plans and thus focuses on medical heterogeneity. Furthermore, we assumed that patients were treated with the appropriate dosage, which is a hard but conservative assumption, because underdosing may dilute our derived effects. In fact, underdosing is common in clinical practice and much more common than overdosing; both should be addressed as a separate problem, regardless of which DOAC is chosen.⁴⁹ Limitations also relate to caveats and challenges before moving to clinical practice, where the magnitude of effects is of central importance.⁴⁵ Our effect sizes were pronounced, considering that 2 highly effective drugs within a substance class were compared. Comparisons between different pharmacologic classes could supposedly have achieved larger (relative) effects. Although significance was comprehensibly expressed on an absolute risk scale, this significance should not be confused with clinical relevance: for clinical impact, it is certainly decisive whether model-assisted decision support can be provided with little effort (automated). If these conditions with a favorable cost-benefit relationship could be met, the projected ARR could translate into quite substantial benefit in clinical practice. However, this estimate of about 1% must be interpreted against the background of no anticoagulant treatment. For an exemplary female patient aged 65 to 74 y with hypertension, diabetes, and a history of myocardial infarction, one can assume an absolute annual risk reduction of 6.2% (rivaroxaban) and 6.8% (apixaban) compared with no anticoagulant treatment (<https://www.sparctool.com>). This is also associated with an absolute annual risk increase for major bleeding of 3.7% (rivaroxaban) and 2.4% (apixaban).

Thus, before such models can be used as bedside tools facilitating decision making, further steps are necessary that are also related to the aforementioned potential limitations: next steps should 1) corroborate the general conclusion by applying the principle to other indications (and situations comparing treatment v. no treatment or different drug classes) with different study designs (robustness), 2) provide a more far-reaching proof-of-

concept in independent data (external validation), and 3) explore the options for presenting risk-based recommendations⁵⁰ in an automated decision-support system (preparatory implementation after clinical utility has been proved in a prospective study). In particular, patients being treated according to such a model's recommendation should be followed up for clinical events in comparison with usual care. Thus, it must be acknowledged that this pioneering work is on the theoretical side and would require further thorough clinical studies before such applications can be confidently used in practice. Even then, the model-assisted recommendations identified here must not overrule clinically binding recommendations (e.g., from the summaries of product characteristics or guidelines).⁵¹ Instead, we have to ensure that such models are used as support systems alongside clinicians, rather than instead of clinicians. Above, one must assess such approaches in general and consider that many medical uncertainties are simply not reducible by information.⁵² One will never ultimately know the truly right treatment for a patient, and decisions will always have to be made by incorporating many probabilities and preferences.⁵³ For this, risk communication is essential⁵³: Supplementary Table S8 shows possible use cases for 4 example patients for whom model-assisted recommendations were derived. The table lists patient characteristics as they enter our models as input characteristics. This results in the respective recommendation for a treatment option, depending on the threshold and weighting of the respective outcome. While the recommendation is presented categorically, it originally stems from a continuous risk scale, so many other visual aids or numerical formats to communicate recommendations could conceivably be used,⁵⁰ including standard formats such as icon arrays or relative formats such as risk ladders.⁵⁴ Moreover, the personalized approach via precision cohorts used here seems promising (in line with the "patients-like-me" framework⁵⁵) and could also be presented to decision makers. Nevertheless, it is important that the patient is closely followed up even after the treatment decision has been made. After a substance has been selected in an adequate dosage, time-varying covariates (e.g., renal function), signs and occurrence of adverse events (e.g., bleeding), interactions with other drugs (e.g., nonsteroidal anti-inflammatory drugs), and adherence in the course of treatment are all expected to modulate responses and should therefore be monitored.

Conclusion



If prescribers are undecided about the potential benefits of alternative therapy options, calculated individual

probabilities for treatment success can support decision making, especially in the case of inconclusive evidence, a differentiated risk-benefit profile, and patients whose complexity deviates from “typical” study populations. Our presented case on the example of DOACs suggested significant ARRs when model-assisted recommendations weighing the risks for stroke and bleeding were considered. There may still be many steps to take on the road to clinical practice, but the principle of decision support developed from routine data can pave the way to future decision-making processes.

Acknowledgements

We thank all participants of the ARMIN study group (ABDA – Federal Union of German Associations of Pharmacists: Christiane Eickhoff, Uta Mueller, Martin Schulz; AOK PLUS: Andreas Fuchs, Dorit Braun, Ulf Maywald; Association of Statutory Health Insurance Physicians–Saxony: Catharina Doehler, Mike Maetzler; Association of Statutory Health Insurance Physicians–Thuringia: Anja Auerbach, Urs Dieter Kuhn, Anke Moeckel; Saxonian Pharmacists’ Association: Christine Hon-scha, Susanne Donner; Thuringian Pharmacists’ Association: Stefan Fink, Kathrin Wagner; Heidelberg University Hospital: Andreas D. Meid, Robert Moecker, Carmen Ruff, Hanna M. Seidling, Felicitas Stoll, Marina Weissenborn, Lucas Wirbka). Andreas D. Meid is thankful for being supported by the Physician-Scientist Programme of the Medical Faculty of Heidelberg University.

ORCID iDs

Andreas D. Meid  <https://orcid.org/0000-0003-3537-3205>
Walter E. Haefeli  <https://orcid.org/0000-0003-0672-6876>

Supplemental Material

Supplementary material for this article is available on the *Medical Decision Making* website at <http://journals.sagepub.com/home/mdm>.

References

1. Armstrong KA, Metlay JP. Annals clinical decision making: communicating risk and engaging patients in shared decision making. *Ann Intern Med.* 2020;172(10):688–92.
2. Meid AD, Ruff C, Wirbka L, et al. Using the causal inference framework to support individualized drug treatment decisions based on observational healthcare data. *Clin Epidemiol.* 2020;12:1223–34.
3. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA.* 2007;298(10):1209–12.
4. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *Int J Epidemiol.* 2016;45(6):2184–93.
5. Armstrong KA, Metlay JP. Annals clinical decision making: translating population evidence to individual patients. *Ann Intern Med.* 2020;172(9):610–6.
6. Chen S, Tian L, Cai T, Yu M. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics.* 2017;73(4):1199–209.
7. Graham DJ, Baro E, Zhang R, et al. Comparative stroke, bleeding, and mortality risks in older Medicare patients treated with oral anticoagulants for nonvalvular atrial fibrillation. *Am J Med.* 2019;132(5):596–604.e11.
8. Lip GYH, Keshishian A, Li X, et al. Effectiveness and safety of oral anticoagulants among nonvalvular atrial fibrillation patients. *Stroke.* 2018;49(12):2933–44.
9. Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical value of predicting individual treatment effects for intensive blood pressure therapy. *Circ Cardiovasc Qual Outcomes.* 2019;12(3):e005010.
10. Yang CY, Lin WA, Su PF, et al. Heterogeneous treatment effects on cardiovascular diseases with dipeptidyl peptidase-4 inhibitors versus sulfonylureas in type 2 diabetes patients. *Clin Pharmacol Ther.* 2021;109(3):772–81.
11. Yao X, Abraham NS, Sangaralingham LR, et al. Effectiveness and safety of dabigatran, rivaroxaban, and apixaban versus warfarin in nonvalvular atrial fibrillation. *J Am Heart Assoc.* 2016;5(6):e003725.
12. Gupta K, Trocio J, Keshishian A, et al. Real-world comparative effectiveness, safety, and health care costs of oral anticoagulants in nonvalvular atrial fibrillation patients in the U.S. Department of Defense population. *J Manag Care Spec Pharm.* 2018;24(11):1116–27.
13. Amin A, Keshishian A, Trocio J, et al. A real-world observational study of hospitalization and health care costs among nonvalvular atrial fibrillation patients prescribed oral anticoagulants in the U.S. Medicare population. *J Manag Care Spec Pharm.* 2018;24(9):911–20.
14. Cha MJ, Choi EK, Han KD, et al. Effectiveness and safety of non-vitamin K antagonist oral anticoagulants in asian patients with atrial fibrillation. *Stroke.* 2017;48(11):3040–8.
15. Marietta M, Banchelli F, Pavesi P, et al. Direct oral anticoagulants vs non-vitamin K antagonist in atrial fibrillation: a prospective, propensity score adjusted cohort study. *Eur J Intern Med.* 2019;62:9–16.
16. Nielsen PB, Skjoth F, Sogaard M, Kjaeldgaard JN, Lip GY, Larsen TB. Effectiveness and safety of reduced dose non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study. *BMJ.* 2017;356:j510.
17. Staerk L, Gerds TA, Lip GYH, et al. Standard and reduced doses of dabigatran, rivaroxaban and apixaban for stroke prevention in atrial fibrillation: a nationwide cohort study. *J Intern Med.* 2018;283(1):45–55.
18. Coleman CI, Peacock WF, Bunz TJ, Alberts MJ. Effectiveness and safety of apixaban, dabigatran, and rivaroxaban versus warfarin in patients with nonvalvular atrial

- fibrillation and previous stroke or transient ischemic attack. *Stroke*. 2017;48(8):2142–9.
19. Yu A, Jeyakumar Y, Wang M, Lee J, Marcucci M, Holbrook A. How personalized are benefit and harm results of randomized trials? A systematic review. *J Clin Epidemiol*. 2020;126:17–25.
 20. Hu L, Ji J, Li F. Estimating heterogeneous survival treatment effect in observational data using machine learning. *Stat Med*. 2021;40:4691–713.
 21. Fang G, Annis IE, Elston-Lafata J, Cykert S. Applying machine learning to predict real-world individual treatment effects: insights from a virtual patient cohort. *J Am Med Inform Assoc*. 2019;26:977–88.
 22. Wendling T, Jung K, Callahan A, Schuler A, Shah NH, Gallego B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Stat Med*. 2018;37:3309–24.
 23. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *J Clin Epidemiol*. 2015;68(2):134–43.
 24. Wirbka L, Haefeli WE, Meid AD. Estimated Thresholds of Minimum Necessary Adherence for Effective Treatment with Direct Oral Anticoagulants - A Retrospective Cohort Study in Health Insurance Claims Data. *Patient Prefer Adherence*. 2021;15:2209–2220.
 25. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43(11):1130–9.
 26. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care*. 2009;47(6):626–33.
 27. 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(2):263–72.
 28. Huling JD, Yu M, Liang M, Smith M. Risk prediction for heterogeneous populations with application to hospital admission prediction. *Biometrics*. 2018;74(2):557–65.
 29. Kunzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci U S A*. 2019;116(10):4156–65.
 30. Vock DM, Wolfson J, Bandyopadhyay S, et al. Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. *J Biomed Inform*. 2016;61:119–31.
 31. Probst P, Wright M, Boulesteix A. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining Knowl Discov*. 2019;9(3):e1301.
 32. Cartus AR, Bodnar LM, Naimi AI. The impact of under-sampling on the predictive performance of logistic regression and machine learning algorithms: a simulation study. *Epidemiology*. 2020;31(5):e42–4.
 33. Dinstag G, Amar D, Ingelsson E, Ashley E, Shamir R. Personalized prediction of adverse heart and kidney events using baseline and longitudinal data from SPRINT and ACCORD. *PLoS One*. 2019;14(8):e0219728.
 34. van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed ‘concordance-statistic for benefit’ provided a useful metric when modeling heterogeneous treatment effects. *J Clin Epidemiol*. 2018;94:59–68.
 35. Sachs MC, Sjolander A, Gabriel EE. Aim for clinical utility, not just predictive accuracy. *Epidemiology*. 2020;31(3):359–64.
 36. Wirbka L, Haefeli WE, Meid AD. A framework to build similarity-based cohorts for personalized treatment advice—a standardized, but flexible workflow with the R package SimBaCo. *PLoS One*. 2020;15(5):e0233686.
 37. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*. 2017;77(1):1–17.
 38. Almutairi AR, Zhou L, Gellad WF, et al. Effectiveness and safety of non-vitamin K antagonist oral anticoagulants for atrial fibrillation and venous thromboembolism: a systematic review and meta-analyses. *Clin Ther*. 2017;39(7):1456–78 e36.
 39. Ruff CT, Giugliano RP, Braunwald E, et al. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. *Lancet*. 2014;383(9921):955–62.
 40. Centers for Disease Control and Prevention. Atrial fibrillation as a contributing cause of death and Medicare hospitalization—United States, 1999. *MMWR Morb Mortal Wkly Rep*. 2003;52(7):128, 130–1.
 41. Kannel WB, Benjamin EJ. Current perceptions of the epidemiology of atrial fibrillation. *Cardiol Clin*. 2009;27(1):13–24, vii.
 42. Li X, Tse VC, Au-Doung LW, Wong ICK, Chan EW. The impact of ischaemic stroke on atrial fibrillation-related healthcare cost: a systematic review. *Europace*. 2017;19(6):937–47.
 43. Wang G, Joo H, Tong X, George MG. Hospital costs associated with atrial fibrillation for patients with ischemic stroke aged 18–64 years in the United States. *Stroke*. 2015;46(5):1314–20.
 44. Qazi JZ, Schnitzer ME, Cote R, Martel MJ, Dorais M, Perreault S. Predicting major bleeding among hospitalized patients using oral anticoagulants for atrial fibrillation after discharge. *PLoS One*. 2021;16(3):e0246691.
 45. Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement. *Ann Intern Med*. 2020;172(1):35–45.
 46. Ruff C, Koukalova L, Haefeli WE, Meid AD. The role of adherence thresholds for development and performance aspects of a prediction model for direct oral anticoagulation adherence. *Front Pharmacol*. 2019;10:113.
 47. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14:137.

48. Steyerberg EW. Validation in prediction research: the waste by data splitting. *J Clin Epidemiol*. 2018;103:131–3.
49. Suwa M, Morii I, Kino M. Rivaroxaban or apixaban for non-valvular atrial fibrillation—efficacy and safety of off-label under-dosing according to plasma concentration. *Circ J*. 2019;83:991–9.
50. Bonner C, Trevena LJ, Gaissmaier W, et al. Current best practice for presenting probabilities in patient decision aids: fundamental principles. *Med Decis Making*. 2021;41(7):821–33.
51. Chen A, Stecker E, Warden BA. Direct oral anticoagulant use: a practical guide to common clinical challenges. *J Am Heart Assoc*. 2020;9(13):e017559.
52. Han PKJ, Strout TD, Gutheil C, et al. How physicians manage medical uncertainty: a qualitative study and conceptual taxonomy. *Med Decis Making*. 2021;41(3):275–91.
53. Han PK, Klein WM, Lehman T, Killam B, Massett H, Freedman AN. Communication of uncertainty regarding individualized cancer risk estimates: effects and influential factors. *Med Decis Making*. 2011;31(2):354–66.
54. Waters EA, Maki J, Liu Y, et al. Risk ladder, table, or bulleted list? Identifying formats that effectively communicate personalized risk and risk reduction information for multiple diseases. *Med Decis Making*. 2021;41(1):74–88.
55. Seligson ND, Warner JL, Dalton WS, et al. Recommendations for patient similarity classes: results of the AMIA 2019 workshop on defining patient similarity. *J Am Med Inform Assoc*. 2020;27(11):1808–12.