



Published in final edited form as:

Clin Pharmacol Ther. 2023 October ; 114(4): 853–861. doi:10.1002/cpt.2983.

Predicting treatment effects of a new-to-market drug in clinical practice based on phase III randomized trial results

HoJin Shin, BPharm, PhD¹, Shirley V. Wang, PhD¹, Dae Hyun Kim, MD, MPH, ScD^{1,2}, Ethan Alt, PhD^{1,*}, Mufaddal Mahesri, MD, MPH¹, Lily G. Bessette, MS¹, Sebastian Schneeweiss, MD, ScD¹, Mehdi Najafzadeh, PhD^{1,**}

¹Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

²Department of Medicine, Division of Gerontology, Beth Israel Deaconess Medical Center, Boston, MA

Abstract

Trial results may not be generalizable to target populations treated in clinical practice with different distributions of baseline characteristics that modify the treatment effect. We used outcome models developed with trial data to predict treatment effects in Medicare populations. We used data from the Randomized Evaluation of Long-Term Anticoagulation Therapy trial (RE-LY), which investigated the effect of dabigatran vs. warfarin on stroke or systemic embolism (stroke/SE) among patients with atrial fibrillation. We developed outcome models by fitting proportional hazards models in trial data. Target populations were trial-eligible Medicare beneficiaries who initiated dabigatran or warfarin in 2010–2011 ('early') and 2010–2017 ('extended'). We predicted 2-year risk ratios (RRs) and risk differences (RDs) for stroke/SE, major bleeding, and all-cause death in the Medicare populations using the observed baseline characteristics. The trial and early target populations had similar mean [SD] CHADS₂ scores (2.15 [1.13] vs. 2.15 [0.91]) but different mean ages (71 vs. 79 years). Compared with RE-LY, the early

Corresponding Author: Dr. HoJin Shin, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street (Suite 3030), Boston, MA 02120. Phone: (617) 278-0930, Fax: (617) 232-8602, hos739@mail.harvard.edu.

* Currently with the Department of Biostatistics, University of North Carolina at Chapel Hill Gillings School of Global Public Health, Chapel Hill, NC, USA

** Currently with Medidata Solutions

Author Contributions

H.S. and S.V.W. wrote the manuscript; H.S., S.V.W., M.N., S.S., D.K., and E.A. designed the research; H.S., S.V.W., M.N., S.S., D.K., and E.A. performed the research; H.S., M.M., and L.B. analyzed the data.

Note: Drs. Najafzadeh and Alt completed this work while employed at Brigham and Women's Hospital. Dr. Najafzadeh is now Senior Director at Medidata Solutions. Dr. Alt is now Assistant Professor at the Department of Biostatistics, University of North Carolina at Chapel Hill Gillings School of Global Public Health, Chapel Hill, NC, USA.

This manuscript is based on research using data from Boehringer Ingelheim that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.

Conflict of Interest statement

Dr. Schneeweiss is principal investigator of the FDA Sentinel Innovation Center funded by the FDA and co-principal investigator of an investigator-initiated grant to the Brigham and Women's Hospital from UCB and Boehringer Ingelheim unrelated to the topic of this study. He is a consultant to Aetion Inc., a software manufacturer of which he owns equity. His interests were declared, reviewed, and approved by the Brigham and Women's Hospital and Partners HealthCare System in accordance with their institutional compliance policies.

Dr. Kim received grants from NIH and personal fees from Alosa Health and VillageMD for unrelated work.

All other authors declared no competing interests for this work.

Medicare population had similar predicted benefit of dabigatran vs. warfarin for stroke/SE (trial RR=0.63;95% CI=0.50 to 0.76 and RD=-1.37%;-1.96% to -0.77%, Medicare RR=0.73;0.65 to 0.82 and RD=-0.92%;-1.26% to -0.59%) and risks for major bleeding and all-cause death. The time-extended target population showed similar results. Outcome model-based prediction facilitates estimating the average treatment effects of a drug in different target populations when treatment and outcome data are unreliable or unavailable. The predicted effects may inform payers' coverage decisions for patients, especially shortly after a drug's launch when observational data are scarce.

Keywords

Generalizability; Trial; RCT; Clinical Practice; Medicare; Anticoagulation; Dabigatran; Warfarin; Outcome risk; Prediction

INTRODUCTION

Randomized controlled trial (RCT) participants with restrictive eligibility criteria may not represent the target population of interest that will be treated in clinical practice.¹ Nonrandomized observational studies could complement the evidence from RCTs with a better representation of the target population. However, results from trials and nonrandomized studies may not be directly comparable because of differences in the distribution of baseline patient characteristics, such as age, sex, or comorbidities, that modify treatment effects.

Outcome model-based risk prediction has been proposed to generalize average treatment effects from RCTs to external target populations by standardizing across multiple potential effect modifiers simultaneously.^{2,3} Briefly, this method transports treatment effects from a trial to target populations by (1) fitting outcome regression models using individual-level data from the trial; (2) predicting probabilities of the outcome based on the distribution of baseline characteristics of the target populations; and (3) averaging the model predictions to generate measures of occurrence (risk or rate) as well as measures of association (relative risk or risk difference). The result is the expected average treatment effect in a target population defined by the observed distribution of baseline patient characteristics. Application of this method could provide valuable insights into the expected benefits and risks of a drug in different target populations, especially shortly after the drug's launch when observational data are scarce.

To illustrate this approach, we developed outcome models leveraging individual-level data from the Randomized Evaluation of Long-Term Anticoagulation Therapy (RE-LY) trial, which investigated the effect of dabigatran vs. warfarin on stroke or systemic embolism among patients with atrial fibrillation (AF).⁴ We then predicted average treatment effects in US Medicare beneficiaries who met the eligibility criteria of RE-LY and compared the predicted vs. observed effects.

METHODS

Study data

Trial—We accessed anonymized data from the RE-LY trial through Vivli (<https://search.vivli.org>). Boehringer Ingelheim was not involved in this decision or any of the conducted analyses. Investigators could request access to the trial data through Vivli's standard data request processes. Analyses of the trial data were exclusively conducted in the Vivli Research Environment, and any results generated from these analyses were subjected to a rigorous evaluation by Vivli to ensure the preservation of patient privacy and data ownership prior to export.

Target populations—We used de-identified data derived from fee-for-service Medicare, a US federal health insurance covering beneficiaries aged ≥65 and aged <65 with disabilities or end-stage renal disease. This database contains individual-level, longitudinal information on demographics, diagnoses and procedures, and outpatient prescription dispensings recorded during billing of all healthcare encounters. The study was approved by the Mass General Brigham Institutional Review Board before data analysis, and licensing agreements were in place. Requests to reproduce findings in our data-analytics environment will be considered.

Study populations, exposure, and eligibility criteria

Trial—The RE-LY trial recruited participants with AF and at least one of the following characteristics (Supplement 1 Table S1): aged ≥75 years with previous stroke or transient ischemic attack, a left ventricular ejection fraction of less than 40%, New York Heart Association class II or higher heart-failure symptoms; or aged 65 to 74 years with diabetes, hypertension, or coronary artery disease. Excluded individuals had a severe heart-valve disorder, a recent stroke, a condition that increased the risk of hemorrhage, renal insufficiency (a creatinine clearance <30 ml per minute), active liver disease, or pregnancy.⁵ Between 2005 and 2007, 18,113 participants were enrolled and 1:1:1 randomly assigned low-dose or high-dose dabigatran (110mg or 150 mg twice daily) or warfarin. We focused on comparing the dabigatran 150 mg and the warfarin treatment arms.

Target populations—We constructed two target populations of interest (Figure 1), one in the early post-market period for dabigatran (October 19, 2010–December 31, 2011), the other encompassing an extended time frame over which there was sufficient accrual of dabigatran-exposed patients to conduct a comparative real-world evidence (RWE) study (October 19, 2010–December 31, 2017). This allowed us to 1) compare the estimated effects of dabigatran vs. warfarin in different target populations over time in clinical practice and 2) compare the predicted effects from generalizing trial results to the observed results in an RWE study that could only be conducted after sufficient exposure data had accrued in clinical practice. Specifically, the early post-market target population included all patients who were eligible for RE-LY and initiated dabigatran or warfarin. The extended time window target population included all patients who were eligible for the trial and were matched on a propensity score predicting the probability of initiating dabigatran as opposed to warfarin. As the observed effects in the RWE study were estimated after matching

on propensity score to adjust for confounding⁶, we generalized effects to the matched population to facilitate comparison of the predicted and observed results in the same population.

Both target populations of interest included RE-LY trial-eligible patients treated with dabigatran or warfarin. We identified those who filled a prescription for dabigatran or warfarin, not simultaneously, with no exposure to warfarin or other DOACs in the previous 183 days; were aged ≥ 65 years; had at least one inpatient or outpatient diagnosis (Supplement 2) of AF in the previous 365 days; had continuous enrollment in Medicare Part A, B, and D and no enrollment in a Medicare Advantage plan in the previous 183 days; and had no missing data on age, sex, race, or geographic region. We then applied other inclusion and exclusion criteria of RE-LY (Supplement 1 Table S1 **for details**). The first dispensing date of the study drugs was defined as cohort entry.

Outcomes and follow-up

Trial—We evaluated three outcomes: stroke or systemic embolism (stroke/SE); major bleeding; and all-cause death. These outcomes were selected because there were sufficient events⁷ for statistical power, and they were measurable in Medicare data. In the RE-LY trial, these outcomes were adjudicated by two independent investigators masked to treatment assignment.⁴ The trial participants were followed based on the intention-to-treat principle.

Target populations—We used only baseline information from the target populations to predict treatment effects. For the RWE study conducted with the extended time window target population, stroke/SE^{8,9,10} and major bleeding^{11,12,13} were identified using validated ICD-9/10-CM diagnosis and procedural code algorithms (Supplement 1 Table S2 and Supplement 2), and all-cause death was ascertained through the Master Beneficiary summary file.¹⁴ We focused on the results from on-treatment analyses because of generally lower treatment adherence in real-world clinical settings compared with RCTs.¹⁵

Predictors and potential effect modifiers of treatment effects

Trial—The candidate predictors and potential effect measure modifiers of the outcomes were limited to the variables available in both the trial and Medicare data. Therefore, vital signs and laboratory data, such as body mass index, systolic and diastolic blood pressures, hemoglobin, and creatinine clearance, could not be used. Additionally, we did not consider variables likely to be severely under-ascertained in claims data, such as smoking status and use of aspirin. Thus, the candidate predictors included age, sex, race, CHADS₂ risk stratification score¹⁶, 10 comorbidities, and 24 concomitant medications (Supplement 1 Table S3 **for the list and coding**).

Target populations—Variables analogous to the candidate predictors available in the RE-LY trial data were defined algorithms based on ICD-9/10-CM procedural and diagnosis codes (Supplement 1 Table S3 and Supplement 2) and prescription dispensing records during the 183 days before or on cohort entry.

STATISTICAL ANALYSIS

Missing data

We excluded one trial participant with multiple variables missing. The RE-LY trial did not provide the exact age of the participants over 89 for privacy reasons. We coded those over 89 years old as 90 to make the age variable continuous. There was no other missing data for candidate predictors considered for outcome models.

Model derivation and validation

In the RE-LY trial data, we fitted Cox proportional hazards models¹⁷ separately within each treatment arm for each outcome (hereafter referred to as ‘dabigatran model’ and ‘warfarin model’) (see Supplement 1 for a stepwise summary and detailed explanation of the model derivation and validation). We selected a Cox proportional hazards model to make the implementation of the outcome model-based risk prediction approach simple and efficient using available R software packages for the relaxed LASSO (least absolute shrinkage and selection operator). The relaxed LASSO method performs variable selection and regularization to reduce the selection of noise variables while maintaining predictive accuracy.^{18,19} This method allowed us to identify the most important predictors of our outcomes while avoiding overfitting the models. We included the CHADS₂ score and age regardless of LASSO selection in each of the models as these are important predictors of the outcomes. For each model, we chose a penalization parameter through 10-fold cross-validation that minimized the model deviance.^{20,21} We evaluated two main domains of model performance²²: discrimination was assessed using Harrell’s C-index ranging from 0.5 (random prediction) to 1 (perfect prediction); and calibration was assessed using a calibration slope (<1 is indicative of overfitting of the model) and a plot comparing the observed Kaplan-Meier estimates against predicted probabilities in different quantiles of the predicted values.²³ Both performance measures were estimated with correction for optimism (hereafter referred to as ‘optimism-corrected’) using bootstrap resampling with 300 iterations^{24,25}, with each being sampled with replacement from the trial data.²⁶ Optimism was calculated by subtracting the mean of bootstrap sample values from the apparent values when the model was fitted to all observations.

Predicting effects of dabigatran vs. warfarin by applying outcome models to generalize results of the RE-LY trial to target populations

We applied the dabigatran and warfarin models to each target population to predict the 2-year probabilities of each of the three outcomes, equivalent to counterfactual probabilities of the outcomes had they been exposed to dabigatran or warfarin. Predicted probabilities were calculated using the cumulative baseline hazard function derived from the Breslow estimator at two years.²⁷ We obtained risks by averaging the predicted probabilities of the outcomes over the covariate distribution for each treatment ($\text{Risk}_{\text{dabigatran}}$ and $\text{Risk}_{\text{warfarin}}$). We then calculated risk ratios ($\text{RRs} = \text{Risk}_{\text{dabigatran}} \div \text{Risk}_{\text{warfarin}}$) and risk differences ($\text{RDs} = \text{Risk}_{\text{dabigatran}} - \text{Risk}_{\text{warfarin}}$). These predicted RRs and RDs respectively quantified the relative and absolute risk differences of dabigatran vs. warfarin had the RE-LY trial been conducted in the target populations. We generalized the intention-to-treat effect estimand for the RE-LY trial, a population that maintained high adherence over study follow-up. Inference was based

on the nonparametric bootstrap resampling.²⁸ We fitted outcome risk prediction models with the LASSO selected predictors in 10,000 bootstrap samples, then applied these models to the target populations to obtain mean RRs and RDs with percentile-based 95% confidence intervals.

Estimating observed effects of dabigatran vs. warfarin in an RWE study among a trial-eligible Medicare population

In the extended time window target population, where there was a sufficient sample size to conduct a comparative RWE study, we used propensity score (PS) matching to adjust for confounding in the estimation of drug effects.⁶ The PS for receiving dabigatran as opposed to warfarin was estimated using a logistic regression model as a function of all prespecified baseline covariates (Supplement 1 Table S4 **for the PS model**). We matched warfarin initiators to dabigatran initiators using 1:1 nearest-neighbor matching without replacement with a caliper of 1% of the PS.^{29,30,31} We compared clinical characteristics between the dabigatran-treated and warfarin-treated groups (Supplement 1 Table S5). A standardized difference <0.1 in baseline covariate distribution between dabigatran and warfarin was considered adequate balance.^{32,33} In the matched cohort, we estimated the 2-year cumulative incidence of outcomes of interest by assuming that the observed rate of events while on treatment would remain constant over two years of follow up (# events*n*2 years/total person-years on treatment), and then estimated RRs and RDs with 95% CIs. The estimand of this study was an on-treatment estimate, with a grace period of 10 days between prescription refills, among those receiving dabigatran or warfarin who were well balanced on all measured potential confounders (Supplement 1 Table S1).

We compared predicted results from the RE-LY trial to the predicted results in the early post-market and the extended time window target populations. We also compared the predicted vs. observed results in the extended time window target population. The magnitude of the differences was quantified through standardized differences (SD_{RR} and SD_{RD}), the difference between treatment effect estimates (natural logarithm for RRs) divided by the standard deviation of the difference.³⁴ Meaningful differences were defined as |standardized difference| >1.96, corresponding to a type 1 error rate of 0.05, which suggests the significance of these differences because only 5% of the repetitions would mistakenly reject the null hypothesis if the study were repeated many times.

All analyses were performed using R version 4.0.2 (<http://www.R-project.org>³⁵; software packages: rms³⁶ and glmnet³⁷) or the Aetion Evidence Platform v4.10 (including R, version 3.4.2 [R Foundation for Statistical Computing]), previously validated.^{38,39,40,41} R source code for model derivation and validation, and prediction of outcome risks are available in the Supplement.

RESULTS

Prediction models

The predictive performance of the outcome models are shown in Figure 2 (Supplement 1 Table S6 **for predictors and coefficients**). Optimism-corrected C-indexes for discrimination

were 0.64 (dabigatran) and 0.65 (warfarin) for stroke/SE, 0.70 and 0.62 for major bleeding, and 0.69 and 0.66 for all-cause death. Optimism-corrected calibration slopes were 0.93 (dabigatran) and 0.94 (warfarin) for stroke/SE, 0.89 and 0.88 for major bleeding, and 0.95 and 0.91 for all-cause death. CHADS₂ score, particularly the combined category of 5 and 6, strongly predicted stroke/SE as expected.⁴²

Demographics

The early post-market target population in clinical practice had a similar CHADS₂ score with smaller estimated variance compared with the RE-LY participants (target population mean=2.15 [standard deviation=0.91] vs. RE-LY mean=2.15 [1.13]) (Table 1 and Supplement 1 Table S7 **for the complete list**). Although older and more likely to be female and white than the RE-LY participants (mean age: target=79 vs. RE-LY=71 years), the early target population had lower prevalence of stroke or transient ischemic attack (5% vs. 20%) and heart failure (13% vs. 32%) and a higher prevalence of cancer (25% vs. 11%). Compared with the early target population, the target population over an extended time frame had overall similar baseline characteristics except for lower prior digoxin use.

Predicted results in RE-LY trial participants compared to predicted results in the early post-market and extended time window target populations in clinical practice

The predicted results for stroke/SE were highly overlapping for the early post-market target and RE-LY populations, indicating similar benefits of dabigatran vs. warfarin in both populations ($SD_{RR}=0.88$, $SD_{RD}=0.78$) (Table 2). However, the modest reductions in risk of major bleeding and all-cause death outcomes observed in RE-LY were not predicted for the early target population in clinical practice, although standardized differences were less than 1.96 for both RR and RD for both major bleeding ($SD_{RR}=1.26$, $SD_{RD}=1.15$) and all-cause death ($SD_{RR}=1.62$, $SD_{RD}=1.72$). The predicted results in the early and extended target populations were similar.

Predicted vs. observed results in the extended time window target population

Compared with the predicted results from generalizing trial data to a target Medicare population in clinical practice identified over an extended time window, the observed results from the RWE study in the same population showed a similar benefit of dabigatran vs. warfarin for stroke/SE ($SD_{RR}=0.67$, $SD_{RD}=0.34$) (Table 2), and greater benefits for major bleeding ($SD_{RR}=3.03$, $SD_{RD}=1.94$) and all-cause death ($SD_{RR}=2.21$, $SD_{RD}=1.68$).

DISCUSSION

Generalization of findings from RCTs to target populations can be problematic in the presence of heterogeneity of treatment effects. We developed outcome models in trial data to demonstrate how to predict the treatment effects of an intervention in target populations treated in clinical practice, accounting for differences in patient characteristics. For the trial-eligible Medicare population in the early post-market period, dabigatran vs. warfarin was predicted to confer a similar benefit for stroke/SE but greater risks for major bleeding and all-cause death compared with the RE-LY trial participants. The predicted benefits and

risks of dabigatran vs. warfarin remained similar in the extended time window Medicare population.

RE-LY trial participants vs. the early post-market target population in clinical practice

The characteristics of the early post-market target population were considerably different from those of RE-LY despite applying the same eligibility criteria to the extent possible in claims data. Notably, the target population was older and more likely to be female and white, with a lower prevalence of risk factors for stroke compared with the trial participants. Nevertheless, the mean CHADS₂ scores were similar in the trial and target populations due to the older population and lower prevalence of CHADS₂ score components (prior stroke or transient ischemic attack and heart failure) offset each other. Consequently, the predicted RRs and RDs for stroke/SE were similar between the trial and target populations because the CHADS₂ score strongly predicted stroke/SE. For major bleeding and all-cause death, the target population was also expected to experience similar risks of dabigatran vs. warfarin, although the point estimates were numerically higher, compared with the trial participants that may be explained by different patient characteristics.

Early post-market vs. extended time window target populations

Compared with 2010–2011, the characteristics of the RE-LY trial-eligible Medicare beneficiaries did not change much over time through 2017 except for lower use of digoxin. Changes in treatment guidelines for AF⁴³ might have influenced medication use. Nevertheless, the predicted benefits and risks of dabigatran vs. warfarin remained similar in the Medicare beneficiaries eligible for RE-LY in 2010–2017. The potential influence of differences in the distributions of effect modifiers between the two target populations on the predicted treatment effects might be modest in magnitude.

Predicted vs. observed treatment effects in the extended time window population

Observed risks were lower than the predicted risks. This may have been because of differences in how the outcome was measured in the trial compared to the RWE study, where in the latter, the outcome algorithms were chosen to have high positive predictive value to minimize bias in the estimation of the RR, sacrificing some sensitivity. While this would tend to reduce bias in the RR, it downwardly biases the magnitude of the RD estimate. The predicted RRs indicated that compared with the RE-LY trial participants, the trial-eligible target population in clinical practice from 2010–2017 was expected to experience a similar benefit for stroke/SE and risks for major bleeding and all-cause death when treated by dabigatran vs. warfarin. Compared with the predicted RRs, the observed ones showed a similar benefit of dabigatran vs. warfarin for stroke/SE but greater benefits for major bleeding and all-cause death. We acknowledge that several factors could contribute to the discrepancies between the predicted and observed results in the extended time window population. These factors may include residual confounding, different methods of outcome ascertainment, and misspecification of the outcome models, the effects of which are challenging to differentiate. Additionally, the comparison between the predicted (intention-to-treat) and observed (on-treatment) results involves different estimands. However, we deemed it a more appropriate comparison compared to comparing the predicted intention-to-treat effect from the trial with the observed intention-to-treat effect in the database study.

These estimates would be expected to differ due to high adherence in trial participants and low adherence in clinical practice.

Strengths

The outcome models developed in trial data provide estimated measures of effect and association based on measured baseline characteristics in a target population. Therefore, residual confounding, different outcome measurement, and shorter duration of follow-up typical in RWE studies are not concerning. Additionally, the outcome model-based prediction requires only one-time model development in a trial, and the same model can be used in any target population that overlaps with the trial population.

Limitations

The application of outcome model-based prediction has several limitations. First, we did not pursue fitting prediction models for rare outcomes, such as intracranial hemorrhage, because the performance of outcome models is greatly affected by the number of events in the trial data.⁴ For rare outcomes, re-weighting^{44,45}, a method generalizing RCT findings to target populations by modeling the probabilities of trial participation, could be used. Another challenge in applying this method was the limited use of covariate information from the trial because only covariates available both in the trial and target populations in the same type, e.g., binary, could be considered candidate predictors. Nevertheless, the calibration of the outcome risk prediction models was decent. Third, the outcome models cannot extrapolate estimates to populations with characteristics explicitly excluded from the trial if these characteristics are deemed potentially crucial to the treatment effect. The RE-LY trial excluded individuals with heart valve disorders or active liver disease potentially associated with the increased risk of stroke or bleeding. Therefore, applying the developed outcome risk prediction models to trial-ineligible target populations warrants caution because the performance of our outcome risk prediction models cannot be guaranteed. Last, predicted treatment effects are affected by measurement errors in the baseline characteristics of the target populations. Under-ascertainment of baseline comorbidities⁴⁶ in our target populations is likely to result in underestimated predicted risks, particularly if the affected variables serve as predictors of the outcomes. Assuming a positive correlation between age and the prevalence of risk factors for stroke, we would have observed higher predicted risks in the target populations.

Conclusions

Outcome model-based prediction facilitates estimating the average treatment effects of a drug in different target populations when treatment and outcome data are unreliable or unavailable. Therefore, the predicted effects may inform payers and health technology assessment agencies' coverage decisions for their patient populations, especially shortly after a drug's launch when observational data are scarce.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Robert J. Glynn* for advice on the manuscript. We thank Dr. Richard Wyss* and Dr. Massimiliano Russo* for their advice on statistical analysis. We also thank Dr. Joshua J. Gagne** for advice on the abstract.

* Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

** Johnson & Johnson, New Brunswick, NJ, USA

Funding information

Dr. Shin is supported by the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School. Dr. Najafzadeh was supported by research grants from the National Institute on Aging (R01AG060163). Dr. Wang was supported by research grants NIA R01AG060163, NHLBI R01HL141505, and NIA R01AG053302. Dr. Schneeweiss was supported by NHLBI R01HL141505 and NIAMS R01AR080194. Dr. Kim is supported by K24AG073527 from NIA.

References

- Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials*. 2015 Nov 3;16:495. doi: 10.1186/s13063-015-1023-4. [PubMed: 26530985]
- Kern HL, Stuart EA, Hill J, Green DP. Assessing methods for generalizing experimental impact estimates to target populations. *J Res Educ Eff*. 2016;9(1):103–127. doi: 10.1080/19345747.2015.1060282. [PubMed: 27668031]
- Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. *Stat Med*. 2020 Jun 30;39(14):1999–2014. doi: 10.1002/sim.8426. [PubMed: 32253789]
- Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, Pogue J, Reilly PA, Themeles E, Varrone J, Wang S, Alings M, Xavier D, Zhu J, Diaz R, Lewis BS, Darius H, Diener HC, Joyner CD, Wallentin L; RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med*. 2009 Sep 17;361(12):1139–51. doi: 10.1056/NEJMoa0905561. [PubMed: 19717844]
- Ezekowitz MD, Connolly S, Parekh A, Reilly PA, Varrone J, Wang S, Oldgren J, Themeles E, Wallentin L, Yusuf S. Rationale and design of RE-LY: randomized evaluation of long-term anticoagulant therapy, warfarin, compared with dabigatran. *Am Heart J*. 2009 May;157(5):805–10, 810.e1-2. doi: 10.1016/j.ahj.2009.02.005. [PubMed: 19376304]
- Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997 Oct 15;127(8 Pt 2):757–63. doi: 10.7326/0003-4819-127-8_part_2-199710151-00064. [PubMed: 9382394]
- Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019 Mar 30;38(7):1276–1296. doi: 10.1002/sim.7992. [PubMed: 30357870]
- Roumie CL, Mitchel E, Gideon PS, Varas-Lorenzo C, Castellsague J, Griffin MR. Validation of ICD-9 codes with a high positive predictive value for incident strokes resulting in hospitalization using Medicaid health data. *Pharmacoepidemiol Drug Saf*. 2008 Jan;17(1):20–6. doi: 10.1002/pds.1518. [PubMed: 17979142]
- Andrade SE, Harrold LR, Tjia J, Cutrona SL, Saczynski JS, Dodd KS, Goldberg RJ, Gurwitz JH. A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data. *Pharmacoepidemiol Drug Saf*. 2012 Jan;21 Suppl 1(Suppl 1):100–28. doi: 10.1002/pds.2312.
- Prat M, Derumeaux H, Sailler L, Lapeyre-Mestre M, Moulis G. Positive predictive values of peripheral arterial and venous thrombosis codes in French hospital database. *Fundam Clin Pharmacol*. 2018 Feb;32(1):108–113. doi: 10.1111/fcp.12326. [PubMed: 29055145]

11. Cunningham A, Stein CM, Chung CP, Daugherty JR, Smalley WE, Ray WA. An automated database case definition for serious bleeding related to oral anticoagulant use. *Pharmacoepidemiol Drug Saf.* 2011 Jun;20(6):560–6. doi: 10.1002/pds.2109. [PubMed: 21387461]
12. Graham DJ, Reichman ME, Wernecke M, Hsueh YH, Izem R, Southworth MR, Wei Y, Liao J, Goulding MR, Mott K, Chillarige Y, MaCurdy TE, Worrall C, Kelman JA. Stroke, Bleeding, and Mortality Risks in Elderly Medicare Beneficiaries Treated With Dabigatran or Rivaroxaban for Nonvalvular Atrial Fibrillation. *JAMA Intern Med.* 2016 Nov 1;176(11):1662–1671. doi: 10.1001/jamainternmed.2016.5954. [PubMed: 27695821]
13. Kim DH, Pawar A, Gagne JJ, Bessette LG, Lee H, Glynn RJ, Schneeweiss S. Frailty and Clinical Outcomes of Direct Oral Anticoagulants Versus Warfarin in Older Adults With Atrial Fibrillation : A Cohort Study. *Ann Intern Med.* 2021 Sep;174(9):1214–1223. doi: 10.7326/M20-7141. [PubMed: 34280330]
14. Research Data Assistance Center (ResDAC). Master Beneficiary Summary File (MBSF) Base. Accessed <https://resdac.org/cms-data/files/mbsf-base> on 31 Oct 2022.
15. Baumgartner PC, Haynes RB, Hersberger KE, Arnet I. A Systematic Review of Medication Adherence Thresholds Dependent of Clinical Outcomes. *Front Pharmacol.* 2018 Nov 20;9:1290. doi: 10.3389/fphar.2018.01290. [PubMed: 30524276]
16. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA.* 2001 Jun 13;285(22):2864–70. doi: 10.1001/jama.285.22.2864. [PubMed: 11401607]
17. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol.* 1972;34:187–202.
18. Tibshirani R Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological).* 1996 Jan;58(1):267–88.
19. Meinshausen N Relaxed lasso. *Computational Statistics & Data Analysis.* 2007 Sep 15;52(1):374–93.
20. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, 2001.
21. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol.* 1995 Dec;48(12):1503–10. doi: 10.1016/0895-4356(95)00048-8. [PubMed: 8543964]
22. Harrell FE. Parametric survival models. In: Harell FE, editor. *Regression Modeling Strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Cham: Springer International Publishing; 2015. p. 423–51. 10.1007/978-3-319-19425-7_18.
23. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010 Jan;21(1):128–38. doi: 10.1097/EDE.0b013e3181c30fb2. [PubMed: 20010215]
24. Booth JG, Sarkar S. Monte Carlo approximation of bootstrap variances. *The American Statistician.* 1998 Nov 1;52(4):354–7. 10.2307/2685441.
25. Efron B Better bootstrap confidence intervals. *Journal of the American statistical Association.* 1987 Mar 1;82(397):171–85. 10.2307/2289144.
26. Steyerberg E (2019) Study design for prediction models. In: Steyerberg E (ed) *Clinical prediction models. a practical approach to development, validation and updating.* Springer, New York, Chapter 5.3, pp 104–109.
27. Breslow N (1972). “Discussion of ‘Regression Models and Life-Tables’ by D.R. Cox” *Journal of the Royal Statistical Society, Series B,* 34(2):216–217.
28. Efron B, Tibshirani RJ. An introduction to the bootstrap. In: *Monographs on Statistics and Applied Probability.* Vol 57. Boca Raton, FL: Chapman & Hall/CRC; 1994.
29. Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *Am J Epidemiol.* 2010 Nov 1;172(9):1092–7. doi: 10.1093/aje/kwq224. [PubMed: 20802241]

30. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011 Mar-Apr;10(2):150–61. doi: 10.1002/pst.433. [PubMed: 20925139]
31. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med*. 2014 Mar 15;33(6):1057–69. doi: 10.1002/sim.6004. [PubMed: 24123228]
32. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009 Nov 10;28(25):3083–107. doi: 10.1002/sim.3697. [PubMed: 19757444]
33. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat Simul Comput*. 2009;38(6):1228–34. doi:10.1080/03610910902859574.
34. Franklin JM, Pawar A, Martin D, et al. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. *Clinical Pharmacology & Therapeutics*. 2020;107(4):817–826. doi:10.1002/cpt.1633. [PubMed: 31541454]
35. RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.
36. Harrell Frank E Jr (2022). rms: Regression Modeling Strategies. R package version 6.3-0. <https://CRAN.R-project.org/package=rms>.
37. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1–22. doi:10.18637/jss.v033.i01. [PubMed: 20808728]
38. Aetion. Evidence Platform[®]. Software for real-world data analysis. Accessed at <http://aetion.com> on 13 January 2022.
39. Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, Bartels DB. Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases. *Clin Pharmacol Ther*. 2016 Mar;99(3):325–32. doi: 10.1002/cpt.329. [PubMed: 26690726]
40. Paterno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using Real-World Data to Predict Findings of an Ongoing Phase IV Cardiovascular Outcome Trial: Cardiovascular Safety of Linagliptin Versus Glimepiride. *Diabetes Care*. 2019 Dec;42(12):2204–2210. doi: 10.2337/dc19-0069. [PubMed: 31239281]
41. Wang SV, Sreedhara SK, Schneeweiss S; REPEAT Initiative. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat Commun*. 2022 Aug 31;13(1):5126. doi: 10.1038/s41467-022-32310-3. [PubMed: 36045130]
42. Fuster V, Rydén LE, Cannom DS, et al. ACC/AHA/ESC 2006 Guidelines for the Management of Patients With Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines (Writing Committee to Revise the 2001 Guidelines for the Management of Patients With Atrial Fibrillation): Developed in Collaboration With the European Heart Rhythm Association and the Heart Rhythm Society. *Circulation*. 2006;114(7). doi:10.1161/CIRCULATIONAHA.106.177292.
43. January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC Jr, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. *J Am Coll Cardiol*. 2014;64(21):e1–76. doi: 10.1016/j.jacc.2014.03.022. [PubMed: 24685669]
44. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of Trial Results Using Inverse Odds of Sampling Weights. *Am J Epidemiol*. 2017 Oct 15;186(8):1010–1014. doi: 10.1093/aje/kwx164. [PubMed: 28535275]
45. Webster-Clark M, Lund JL, Stürmer T, Poole C, Simpson RJ, Edwards JK. Reweighting Oranges to Apples: Transported RE-LY Trial Versus Nonexperimental Effect Estimates of Anticoagulation in Atrial Fibrillation. *Epidemiology*. 2020 Sep;31(5):605–613. doi: 10.1097/EDE.0000000000001230. [PubMed: 32740469]

46. Kern EF, Maney M, Miller DR, Tseng CL, Tiwari A, Rajan M, Aron D, Pogach L. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv Res.* 2006 Apr;41(2):564–80. doi: 10.1111/j.1475-6773.2005.00482.x. [PubMed: 16584465]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Study Highlights

What is the current knowledge on the topic?

Trial results may not be generalizable to a target population treated in clinical practice with a different distribution of baseline characteristics, such as age or comorbidities, that may modify the treatment effect.

What question did this study address?

We leveraged individual-level RE-LY trial data to predict average treatment effects in older and frail Medicare populations treated in clinical practice.

What does this study add to our knowledge?

Medicare populations who met the eligibility criteria of RE-LY were predicted to experience similar benefits for stroke or systemic embolism, major bleeding, and all-cause death when treated with dabigatran vs. warfarin compared with RE-LY participants.

How might this change drug discovery, development, and/or therapeutics?

Predicted treatment effects in a target population provide valuable insights into the expected benefits and risks, especially shortly after a drug's launch when observational data are scarce. These predicted effects could be informative for payers to make coverage decisions for patients.

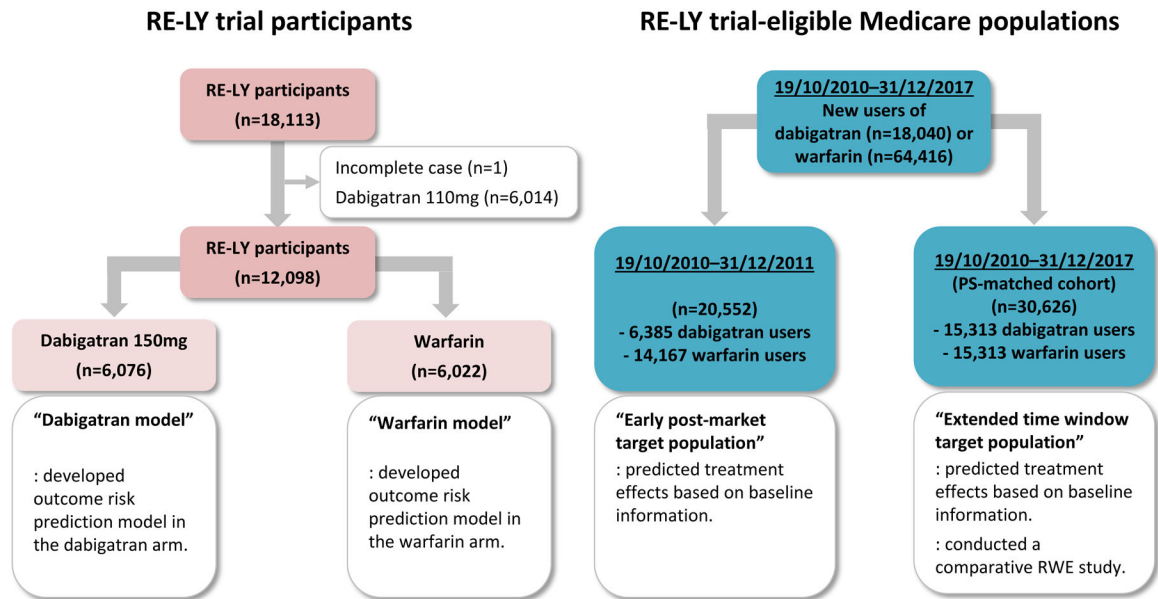


Figure 1. Study Implementation
No legend

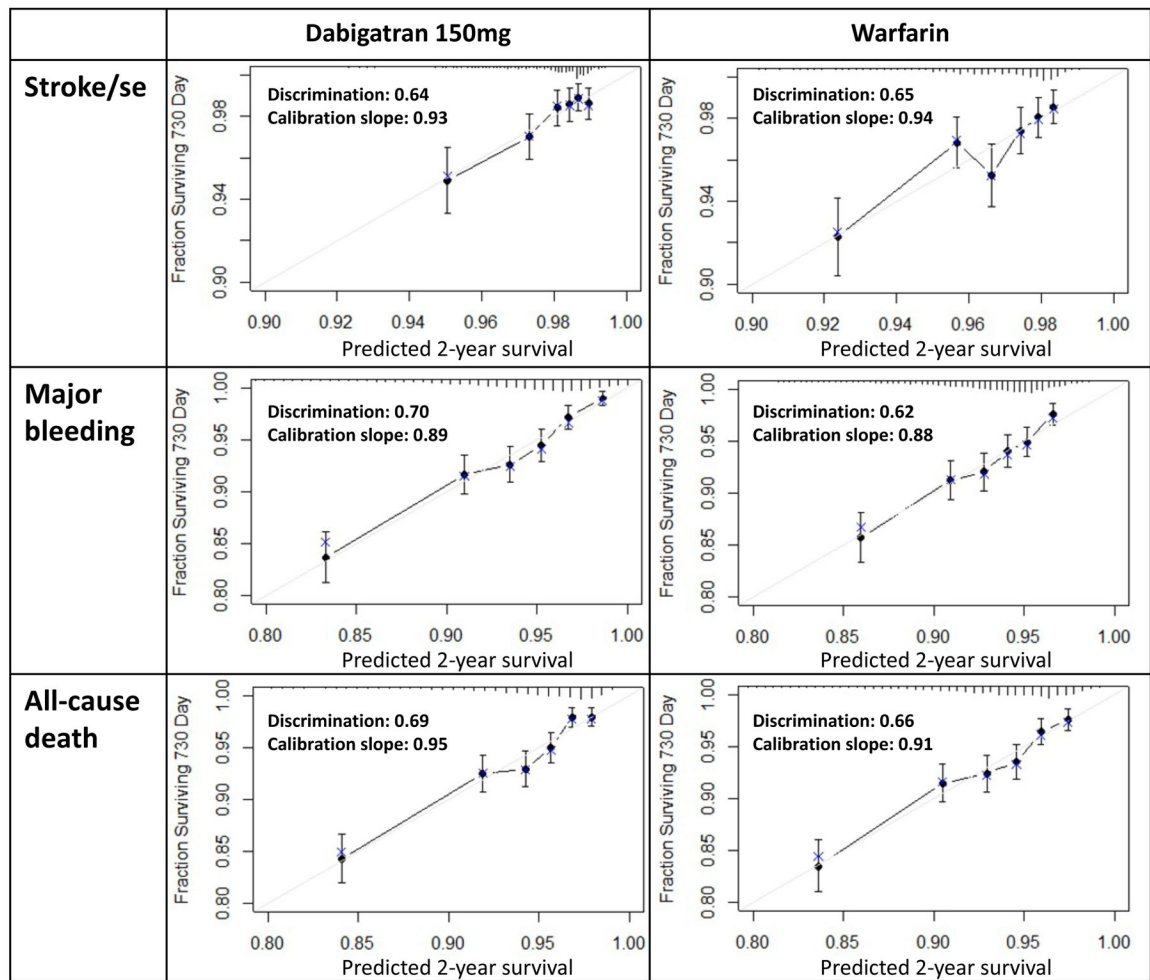


Figure 2. Prediction model performance

Note: discrimination and calibration slope were optimism-corrected values through bootstrap resampling.

Gray diagonal line: ideal line for a slope of 1.

Black dot: in-sample calibration accuracy.

Blue x: bootstrap estimates corrected for overfitting.

Table 1.

Demographics, n (%) unless otherwise specified.

Baseline characteristics	RE-LY trial	Medicare	
	dabigatran 150mg + warfarin (n=12,098)	Early post-market target population (n=20,552)	Extended time window target population (n=30,626)
Demographics			
Age; mean (sd)	71.4 (8.6)	78.6 (6.4)	77.6 (6.23)
<65	1983 (16.4)	0 (0.0)	0 (0.0)
>=65-<75	5226 (43.2)	4,874 (23.7)	8,366 (27.3)
>=75	4889 (40.4)	15678 (76.3)	22260 (72.7)
Sex (Female)	4449 (36.8)	9948 (48.4)	13772 (45.0)
Race (White)	8471 (70.0)	19526 (95.0)	29013 (94.7)
Risk stratification score for AF			
CHADS2 score; mean (sd)	2.15 (1.13)	2.15 (0.91)	2.11 (0.91)
0 or 1	3823 (31.6)	4282 (20.8)	6809 (22.2)
2	4365 (36.1)	10465 (50.9)	15667 (51.2)
3	2420 (20.0)	4298 (20.9)	5964 (19.5)
4	1071 (8.9)	1201 (5.8)	1769 (5.8)
5 or 6	419 (3.5)	306 (1.5)	417 (1.4)
Comorbidities			
Stroke or TIA	2428 (20.1)	999 (4.9)	1949 (6.4)
Myocardial infarction	1997 (16.5)	898 (4.4)	1824 (6.0)
Heart failure	3856 (31.9)	2587 (12.6)	3466 (11.3)
Diabetes	2812 (23.2)	5587 (27.2)	8190 (26.7)
Coronary artery disease	3373 (27.9)	6174 (30.0)	9254 (30.2)
Hypertension	9545 (78.9)	17062 (83.0)	25529 (83.4)
Cancer	1272 (10.5)	5169 (25.2)	7468 (24.4)
Medications in use at baseline			
ACE Inhibitor	5424 (44.8)	7787 (37.9)	11184 (36.5)
ARB	2888 (23.9)	4193 (20.4)	6962 (22.7)
Amiodarone	1329 (11.0)	1521 (7.4)	2399 (7.8)
Beta blocker	7609 (62.9)	13352 (65.0)	20428 (66.7)
Clopidogrel	682 (5.6)	2047 (10.0)	3430 (11.2)
Digoxin	3509 (29.0)	3564 (17.3)	3218 (10.5)
Diltiazem	1120 (9.3)	3356 (16.3)	5060 (16.5)
Diuretic	6189 (51.2)	5230 (25.4)	6317 (20.6)
H ₂ -receptor antagonist	519 (4.3)	719 (3.5)	1234 (4.0)
Oral Hypoglycemic	2004 (16.6)	4195 (20.4)	6332 (20.7)
Other NSAID	613 (5.1)	2116 (10.3)	3866 (12.6)

	RE-LY trial	Medicare	
Baseline characteristics	dabigatran 150mg + warfarin (n=12,098)	Early post-market target population (n=20,552)	Extended time window target population (n=30,626)
Other antiarrhythmic	981 (8.1)	3105 (15.1)	4619 (15.1)
Parenteral anticoagulant	642 (5.3)	456 (2.2)	676 (2.2)
Proton-pump inhibitors	1720 (14.2)	3329 (16.2)	5544 (18.1)
Statin	5355 (44.3)	11159 (54.3)	17049 (55.7)
Any antithrombotic treatment	5378 (44.5)	2595 (12.6)	4219 (13.8)

Abbreviations: ACE, angiotensin-converting enzyme; ARB, angiotensin II receptor blocker; COX, cyclooxygenase; H₂-receptor: histamine type-2 receptor; NSAID, nonsteroidal anti-inflammatory drug; RE-LY, Randomized Evaluation of Long-Term Anticoagulation Therapy; sd, standard deviation; TIA, transient ischemic attack; VKA, vitamin K antagonist.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Predicted and observed results in RE-LY and Medicare populations

Outcome	Population*	Two-year Risk (%)				RR (95% CI)	RD (95% CI)	SD _{RD} [†]
		Dabigatran	Warfarin	(ref: warfarin)	SD _{RR} [‡]			
Stroke/SE	Predicted	RE-LY	2.22	3.59	0.62 (0.49, 0.78)		-1.37 (-2.04, -0.72)	
		Early	2.48	3.39	0.74 (0.53, 1.00)	0.88	-0.92 (-1.86, -0.01)	0.78
		Extended	2.37	3.27	0.73 (0.53, 0.98)	0.83	-0.90 (-1.76, -0.06)	0.86
	Observed		1.95	3.01	0.65 (0.56, 0.75)	0.67	-1.06 (-1.41, -0.71)	0.34
Major bleeding	Predicted	RE-LY	6.67	7.42	0.90 (0.75, 1.05)		-0.75 (-1.91, 0.34)	
		Early	8.82	8.24	1.08 (0.85, 1.34)	1.26	0.58 (-1.39, 2.56)	1.15
		Extended	8.43	7.81	1.09 (0.86, 1.35)	1.33	0.62 (-1.24, 2.50)	1.23
	Observed		3.61	4.89	0.74 (0.66, 0.82)	3.03	-1.28 (-1.73, -0.82)	1.94
All-cause death	Predicted	RE-LY	6.56	7.51	0.88 (0.76, 1.01)		-0.95 (-1.93, 0.04)	
		Early	6.64	6.01	1.11 (0.87, 1.41)	1.62	0.64 (-0.86, 2.17)	1.72
		Extended	5.99	5.44	1.11 (0.87, 1.39)	1.66	0.55 (-0.80, 1.92)	1.75
	Observed		3.81	4.62	0.83 (0.74, 0.92)	2.21	-0.81 (-1.26, -0.36)	1.68

Abbreviations: RR, risk ratio; RD, risk difference; SD, standardized difference.

* Early: early post-market target population in 2010–2011; Extended: extended time window target population in 2010–2017; RE-LY: the RE-LY trial target population.

[†] Relative to the predicted results in the RE-LY trial participants except for the observed results in the extended time window target population that are relative to the predicted results in the same extended time window target population.