Simulation for Predicting Effectiveness and Safety of New Cardiovascular Drugs in Routine Care Populations

Mehdi Najafzadeh[1], Sebastian Schneeweiss[1], Niteesh K. Choudhry[1], Shirley V. Wang[1] and Joshua J. Gagne[1]

In the presence of heterogeneity of treatment effect (HTE), the average treatment effect from a randomized controlled trial (RCT) may not be applicable to different patients, such as those in observational settings. Our objective was to develop a novel approach that uses individual-level simulation to expand RCT results to target patient populations in the presence of HTE. For this purpose, we compared the results of the Randomized Evaluation of Long-Term Anticoagulation Therapy (RE-LY) trial, and two observational studies that compared benefits and risks of dabigatran to warfarin in patients with atrial fibrillation. We developed a simulation model that replicates the rates of ischemic stroke and major bleeding observed in RE-LY using published outcome risk models and participants' baseline characteristics. We used our validated simulation model to predict what the results of the RCT would have been had it been conducted in populations similar to those in the observational studies.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ The results of observational studies often differ from those of RCTs. One possible explanation is that in the presence of heterogeneity of treatment effect (HTE), the average treatment effect from a randomized clinical trial (RCT) may not be applicable to different patients, such as those in observational studies.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Our objective was to quantify the differences between RCT and observational studies that can be explained by the differences in study populations.

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE

☑ We proposed a novel approach that uses individual-level simulation to predict what the results of the RCT would have been had it been conducted in populations similar to those in the observational studies.

HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE

☑ Application of this method can facilitate interpretation of between-study variations in results and can support deriving better informed inferences about true underlying treatment effect.

Randomized controlled clinical trials (RCTs) are usually considered the gold standard for assessing treatment efficacy.[1] Random assignment of patients to different treatments can yield unbiased estimates of treatment effect for the population under study. Observational studies can complement RCTs by capturing rare adverse events and long-term outcomes that are critical for informing patient-centered treatment decisions, but are not usually available in RCTs.[1] Observational studies can also provide outcome estimates of treatment effectiveness in broad patient populations in actual practice settings, rather than in narrow RCT populations in protocol-directed experimental settings.

The results of observational studies often differ from those of RCTs.[2,3] Such discrepancies can arise from several sources, including confounding bias in the observational study, differences in outcome measurement, differences in follow-up, differences in adherence, and from differences in patient characteristics (**Table 1**) that modify the treatment effect, known as heterogeneity of treatment effect (HTE). In the absence of HTE, results of RCTs should, in expectation, generalize to other patient populations provided that doses, adherence, and durations of treatment are similar.[4] However, in the presence of HTE the average treatment effect from the RCT may not be applicable to many patients who use the treatments in typical care settings.[5]

We propose a novel approach to expand RCT results to broader patient populations in the presence of HTE. The approach uses an individual-level simulation and evidence about

Correction added on March 29, 2018, after first online publication: In the author list, the name "Sebastian S. Schneeweiss" was changed to "Sebastian Schneeweiss".

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. Correspondence: Mehdi Najafzadeh (mnajafzadeh@bwh.harvard.edu)

Received 8 September 2017; accepted 28 January 2018; advance online publication 00 Month 2018. doi:10.1002/cpt.1045

Table 1 Baseline characteristics of patients in RE-LY and two observational studies

Baseline characteristics	RE-LY (2009)	Graham (2014)	Larsen (2013)
Age (mean, SD)	71.5 (8.8)	Estimated ~79	67.4 (8.5)
65-74		42%	
75-84		43%	
85+		16%	
Male	63.2%	49%	61.5%
T2D	23.1%	33%	12.1%
Hypertension	78.9%	87%	22.7%
Heart failure	31.8%	18%	5.2%
Stroke/TIA	20.3%	23% (=Stroke+TIA+other cerebrovascular)[a]	17.1%
CAD	31.8% (=HF)[a]	48% (=Other ischemic heart disease)[a]	5.2% (=HF)[a]
ASA co-med	38.7%	38.7% (=RE-LY)[a]	35.2%
Clopidogrel co-med	16.9 % (=MI)[a]	17%	5%
MI	16.9%	18% (=MI+coronary revascularization)[a]	6.1
CHADS2 (mean, SD)	2.2 (1.2)	2.7 (1.25) (estimated)	0.96 (1.07)
0-1	32.2%	28%	
2	35.2%	40%	
3	32.6% (for 3+)	21%	9.5% (for 3+)
4+		10%	

CHADS2 = 1*HF + 1*Hypertension + 1*T2D + 1*Age>75 + 2*Stroke/TIA.
[a]Missing values for these variables were imputed based on other variables that had been directly reported in the study. For example, % of patients with history of stroke/TIA in Graham *et al*. study was approximated by adding % of patients with history of stroke, TIA, and other cerebrovascular diseases.

HTE derived from the RCT to estimate effects applicable to the characteristics of external patient populations.

RESULTS

Replication of RCT results

The simulation model (**Supplemental Figure S1**) successfully replicated the overall results of RE-LY (**Table 2**). The simulation model predicted rates of stroke/SE of 1.13 and 1.74, ischemic stroke of 0.94 and 1.23, ICH of 0.32 and 0.76, and major bleeding of 3.34 and 3.58 per 100 person-years for dabigatran and warfarin groups, respectively. These estimates closely matched the observed rates in RE-LY. The hazard ratios (HRs) and 95% credible intervals (CIs) obtained from the simulation model also matched the HRs from RE-LY (**Table 2**). Modeling error, defined as the difference between observed and predicted results of RE-LY, could have been caused by misspecification of simulation model structure or assumptions about input parameters, especially misclassification of risk strata based on CHADS2 subgroups.

Counterfactual results of RE-LY in the Graham 2014 population

We generated cohorts of equal size and with similar covariate distributions as those in the Graham 2014 study.[20] The simulation model predicted rates of ischemic stroke of 1.13 and 1.45 (HR, 0.78; 95% CI, 0.61–1.02), rates of ICH of 0.37 and 0.83 (HR, 0.45; 95% CI, 0.26–0.69), and rates of major bleeding of 3.85 and 3.92 per 100 person-years (HR, 0.98; 95% CI, 0.83–1.15) in dabigatran and warfarin groups, respectively (**Table 2**). These results suggested that the slightly higher rates of ischemic stroke (1.13 and 1.39 in dabigatran and warfarin, respectively) and ICH (0.33 and 0.96 in dabigatran and warfarin, respectively) observed in Graham 2014 compared to RE-LY can be almost entirely explained by differences in patients' baseline characteristics. Rates of stroke/SE were not reported in the Graham *et al*. study, but were predicted to be 1.37 and 2.04 in dabigatran and warfarin groups (HR, 0.67; 95% CI, 0.52–0.87), respectively, using our simulation model.

The model predicted higher rates of major bleeding in the Graham 2014 population (3.85 and 3.92 in dabigatran and warfarin, respectively) compared to the RE-LY population. Predicted rates were smaller than corresponding observed rates in Graham 2014 (4.27 and 4.39 in dabigatran and warfarin, respectively). This suggests that differences in rates of major bleeding between RE-LY and Graham *et al*. could be only partially attributed to differences in study populations. However, the predicted (HR, 0.98; 95% CI, 0.83–1.15) and observed (HR, 0.97; 95% CI, 0.88–1.07) HRs for major bleeding were similar.

We also examined discrepancies in the difference measure scale between the results of Graham 2014 and RE-LY by comparing risk differences in ischemic stroke, ICH, and major bleeding. We determined the contribution of HTE, modeling error, and

Table 2 Observed vs. predicted rates and hazard ratios in RE-LY and two observational studies

Observed results	RE-LY			Graham *et al.*			Larsen *et al.*		
	Dabi150	Warfarin	HR (95% CrI)	Dabi150	Warfarin	HR (95% CrI)	Dabi150	Warfarin	HR (95% CrI)
Stroke/SE	1.11	1.69	0.66 (0.53–0.82)						
Ischemic stroke	0.92	1.20	0.76 (0.60–0.98)	1.13	1.39	0.8 (0.67–0.96)	3.5	3.0	1.18 (0.85–1.64)
ICH	0.3	0.74	0.4 (0.27–0.60)	0.33	0.96	0.34 (0.26–0.46)	0.1	0.7	0.08 (0.01–0.40)
Major Bleeding	3.11	3.36	0.93 (0.81–1.07)	4.27	4.39	0.97 (0.88–1.07)	2.2	2.9	0.77 (0.51–1.13)
Predicted results	**RE-LY**			**Graham *et al.***			**Larsen *et al.***		
	Dabi150	Warfarin	HR	Dabi150	Warfarin	HR	Dabi150	Warfarin	HR
Stroke/SE[a]	1.13 (0.87–1.42)	1.74 (1.40–2.08)	0.65 (0.47–0.86)	1.37 (1.08–1.67)	2.04 (1.68–2.40)	0.67 (0.52–0.87)	0.79 (0.57–1.02)	1.27 (1.00–1.57)	0.62 (0.43–0.89)
Ischemic stroke	0.94 (0.72–1.17)	1.23 (0.99–1.48)	0.76 (0.55–1.00)	1.13 (0.90–1.38)	1.45 (1.20–1.70)	0.78 (0.61–1.02)	0.66 (0.47–0.84)	0.91 (0.71–1.11)	0.72 (0.50–1.04)
ICH	0.32 (0.18–0.48)	0.76 (0.55–0.97)	0.42 (0.22–0.62)	0.37 (0.23–0.55)	0.83 (0.63–1.07)	0.45 (0.26–0.69)	0.24 (0.13–0.38)	0.66 (0.47–0.88)	0.37 (0.18–0.63)
Major bleeding	3.34 (2.92–3.82)	3.58 (3.13–4.03)	0.93 (0.79–1.11)	3.85 (3.37–4.33)	3.92 (3.47–4.40)	0.98 (0.83–1.15)	2.51 (2.13–2.92)	3.06 (2.63–3.47)	0.82 (0.67–1.01)

CrI, credible interval.
[a]The numbers inside parentheses are Monte Carlo confidence intervals assuming 6,000 simulated patients in each treatment arm.

unexplained discrepancies to explain these between study differences (**Figure 1**). For example, **Figure 1a** indicates how HTE, modeling error, and unexplained components add up to explain the discrepancy between the observed risk difference of ischemic stroke in Graham 2014 (RD, –0.26) and the observed risk difference in RE-LY (RD, –0.28). Note that the sum of HTE, modeling error, and unexplained component for a given outcome equals the discrepancy between the observed risk difference in RE-LY and the observed risk difference for that outcome in Graham 2014. **Figure 1c** also suggests that a substantial amount of discrepancy in major bleeding results between Graham 2014 and RE-LY can be explained by HTE. **Figure S2** demonstrates discrepancies between predicted and observed estimates of risk difference in ischemic stroke and major bleeding in a benefit–risk plane. We also present the joint distribution of predicted risk differences in **Figure S3**.

Counterfactual results of RE-LY in the Larsen 2013 population

When we changed the baseline characteristics of the simulated cohort to match those in Larsen 2013[21] and repeated the simulation, predicted rates of ischemic stroke were 0.66 and 0.91 per 100 person-years for dabigatran and warfarin groups, respectively (HR, 0.72; 95% CI, 0.50–1.04) (**Table 2**). In contrast, observed rates of ischemic stroke were 3.5 and 3.0 (HR, 1.18; 95% CI, 0.85–1.64) in Larsen 2013. The predicted ICH rates were 0.66 and 0.91, respectively (HR, 0.72; 95% CI, 0.50–1.04) in a population with characteristics similar to those in Larsen 2013, whereas the observed rates were 0.1 and 0.7, respectively (HR, 0.08; 95% CI, 0.01–0.40).

The predicted rates of major bleeding were 2.51 and 3.06, respectively (HR, 0.82; 95% CI, 0.67–1.01) in a population similar to the Larsen 2013. These predicted rates were slightly larger but comparable to observed rates of 2.2 and 2.9 (HR, 0.77; 95% CI, 0.51–1.13). **Figure 1a** indicates how HTE, modeling error, and unexplained components add up to explain the discrepancy between the observed ischemic stroke risk difference in Larsen 2013 (RD, 0.5) and in RE-LY (RD, –0.28). **Figure 1a** suggests that the discrepancy between the observed risk difference for ischemic stroke in Larsen 2013 and RE-LY is mainly due to an unexplained component. **Figure 1c** suggests that a substantial portion of the discrepancy between the observed Larsen 2013 results and the RE-LY results is due to HTE.

DISCUSSION

Observational studies using databases of routinely collected health information can be useful for examining whether the benefits and risks of treatments found in RCTs apply also to populations of patients treated in routine care settings. Paradoxically, however, doubt about the validity of the observational studies often arises when they yield results that are discrepant with randomized investigations. These comparisons, however, often ignore between-study population differences and the impact that they can have on the results in the presence of HTE. We propose a simulation method to predict the counterfactual outcomes of a randomized trial had it been conducted in a target observational study population. We used the approach to estimate predicted outcomes of the RE-LY trial in populations similar to two observational studies and quantified how much of the differences in results could be

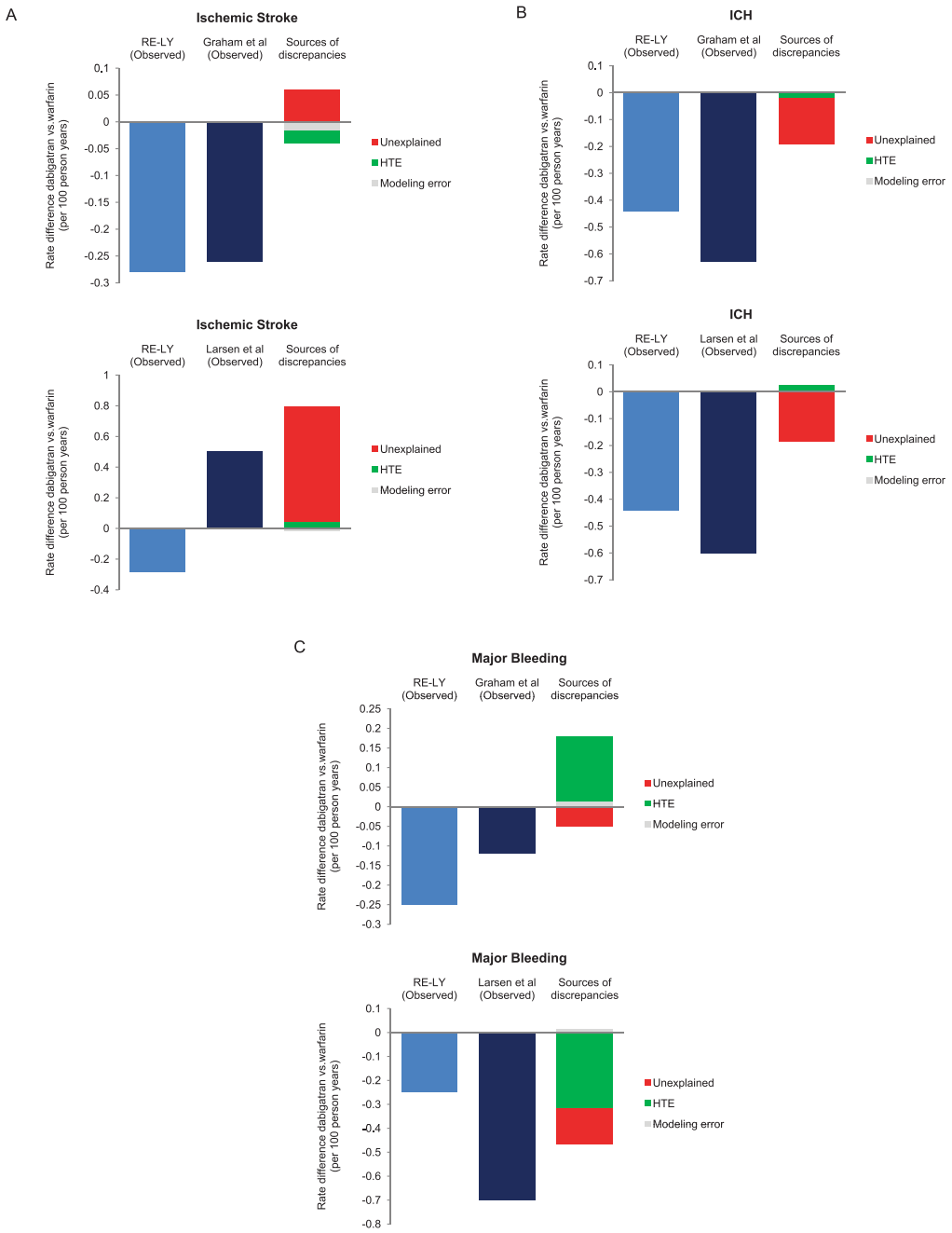

Figure 1 Risk differences observed in RE-LY vs. those observed in Graham 2014 and Larsen 2013 and contribution of HTE and simulation modeling error to observed between study discrepancies. Note that in each panel the algebraic sum of modeling error, HTE, and unexplained components is equal to discrepancies of observed risk differences between RE-LY and Graham 2014 and Larsen 2013. Magnitude of modeling error, HTE, and unexplained components are based on the definitions provided in Table S4. More specifically, modeling error was defined as the difference between observed and predicted RCT outcomes; HTE was the difference between predicted outcomes of RCT and predicted outcomes of observational study; and unexplained component was defined as the difference between predicted and observed outcomes of observational study. **(a)** Ischemic stroke. **(b)** Intracranial hemorrhage (ICH). **(c)** Major bleeding. [Color figure can be viewed at cpt-journal.com]

explained by HTE as a result of different study populations vs. other differences between the studies, such as bias or differences in outcome definitions, adherence, or follow-up duration.

Our results suggest that differences in the rates of ischemic stroke, ICH, and major bleeding between the RE-LY trial and Graham 2014, were largely explained by differences in study populations. However, observed differences between rates of stroke and ICH outcomes in RE-LY and Larsen 2013 could not be explained by differences in study populations. In particular, the difference in rates of ischemic stroke between the RE-LY trial and Larsen 2013 became larger after accounting for differences in study populations. This suggested

that factors other than study population differences, such as confounding or misclassification of outcomes, likely explain the differences in rates of outcomes.[6]

We also examined discrepancies in treatment effect sizes between studies by comparing rate differences for ischemic stroke, ICH, and major bleeding. Our results suggest that discrepancies in ischemic stroke rate differences between Larsen 2013 and the RE-LY trial remained after adjusting for differences between study populations, indicating that the observed differences are likely due to factors other than HTE. However, discrepancies in rate differences for major bleeding could be largely explained by HTE due to differences in study populations.

Although many factors can contribute to differences in between-study results, here we focused on differences resulting from HTE and variation in study populations. Study populations are rarely the same across different studies. In most situations, baseline covariates (e.g., age, history of stroke) are also risk factors for the outcomes that are being studied and, therefore, they can differentially influence outcome event rates across different patient populations. In the presence of HTE, baseline covariates can influence the treatment effect size measured in different patients. Therefore, differences in study populations can result in variation in estimated rates and treatment effect sizes across different studies. Our proposed method provides an approximate approach to estimate the magnitude of this impact on study results. Selection bias, information bias, and confounding remain challenging issues in observational studies in routinely collected databases. Although by using methods such as propensity scoring or stratification we can reduce measured confounding in observational studies, residual confounding due to unmeasured covariates may not be fully addressed by using these methods. Information bias, which is caused by misclassification of outcomes, exposures, or covariates, can also distort study results, especially in observational studies where adjudication of outcomes are not feasible. Sampling variation (chance), even if the samples were drawn from the same patient population, could also result in between study differences. However, the variation in the point estimates due to random sampling is expected to be captured in the reported confidence intervals.

To our knowledge, this is the first study that used simulation to predict counterfactual outcomes of a randomized trial that are applicable to a different target population in the presence of heterogeneity of treatment effect. This simulation method can be applied in the context of any other treatment or health condition. Cole and Stuart[7] proposed using inverse probability-of-selection weights as a method to generalize inferences from a randomized trial to a specified target population. Using this method they demonstrated that the effect of an HIV treatment from a randomized trial would be 12% smaller in the target population of US people infected with HIV in 2006. The method reweights the randomized trial results to provide standardized estimates of treatment effect in a target population. An important practical limitation of this method and of related standardization approaches[8] (e.g., standardized mortality ratios), which use strata-specific estimates and combines them using a weighted average based on covariate distributions of a target population, is that it requires knowing the joint distribution of effect modifiers in the target population, which may not always be available or obtainable. In contrast, our proposed approach provides an alternative method that builds on outcome prediction models to project a wide range of outcomes and measures such as rates, relative risks, and risk differences in a target population. This is particularly useful when the goal is between study comparison of rates for multiple outcomes, treatment effect sizes, and decomposing sources of discrepant results. Basu *et al*.[9] estimated counterfactual distribution of blood pressure in a certain population (e.g., blacks) if its risk factor profile (e.g., body mass index, sodium intake, and alcohol use) looked like a comparator population (blacks). Their approach was based on the Oaxaca-Blinder decomposition method that has been used to decompose sources of difference in outcomes in different populations using regression analysis.[10] Our work is also closely related to the recent literature about transportability in epidemiology. Methods such as do-calculus developed in this area aim to assess external validity and applicability of experimental results in different target populations. In particular, methods such as do-calculus provide formal discussion about the necessary conditions under those transporting experimental results to a target population is valid.

We presented our proposed approach by showing how an individual-level simulation and evidence about heterogeneity of treatment effect can be derived from a randomized trial to estimate effects applicable to the characteristics of "real-world" patient populations. This may be especially useful in the early phase after drug approval when physicians, decision makers, and payors speculate about the extent to which treatment effect is applicable to different target populations. Application of the proposed method can also be generalized to adjust between-study population differences when comparing two or more observational studies as long as an outcome prediction model is developed and validated in one of the studies.

Several limitations of the proposed method as well as the particular case studies used should be mentioned. The accuracy of the model was limited by lack of availability of multivariable outcome prediction models. Specifically, in the absence of individual-level trial data, the model relied on the results of an analysis that reported only risk of outcomes based on three levels of patients' CHADS2 scores. Having access to patient-level RCT data may have allowed us to develop and validate more accurate outcome prediction models. Reilley *et al*. conducted a follow-on analysis of RE-LY data to predict risk of ischemic stroke/SE and major bleeding based on a number of patient characteristics.[11] Although their models were more complex, they were only reported for the dabigatran arm, precluding us from using it to estimate outcome rates among both dabigatran and warfarin users. However, we conducted an *ad hoc* analysis using the prediction models reported in Reilley *et al*. to evaluate our simulation model for dabigatran users (**Table S1**). As with our primary model, this alternative model also successfully replicated the dabigatran outcome rates in RE-LY. Also in line with the results of our primary analysis, these predicted rates suggested that discrepancies in the rates of ischemic stroke between RE-LY and Larsen *et al*. could not be explained based on differences in study

population, and are most likely indicative of bias in the observational study. Potential errors in measurement of covariates and outcomes using observational data is another limitation in implementing this approach. Unlike RCTs, covariates and outcomes in observational studies are captured using data that are not typically recorded for research purposes.[12] For example, outcomes and covariates in Graham 2014 were derived from Medicare billing information, while Larsen 2013 used hospital records. Therefore, some of the unexplained differences, particularly between RE-LY and Larsen 2013, could be due to variation in measurement of covariates and outcomes. One possible solution to this problem can be achieved by linking RCTs with observational studies. This type of linkage would create "information overlap" that could be used for calibrating measurement using observational data sources.[12]

Our approach is based on the assumption that the outcome prediction models developed based on RCT data can correctly and fully describe the relationships between covariates, treatment, and outcomes. If the prediction model is misspecified, such as by omitting effect modifiers or interactions among variables that led to HTE, the simulated outcomes in the target cohort may differ from those of the RCT. Developing outcome prediction models requires having access to either outcome prediction models or to patient-level RCT data to estimate such statistical models. Nascent initiatives encouraging open access to individual-level RCT data[13,14] can facilitate developing more precise simulation models and, therefore, wider application of this method for understanding HTE between studies and populations. Dahabreh *et al*.[15] provide a concise discussion about benefits of using multivariate outcome prediction models rather than conventional "one-variable-at-a-time" subgroup analyses. Studies conducted by Burke *et al*.,[16] Kent *et al*.,[17] and Reilley *et al*.[11] demonstrate how multivariate outcomes prediction models based on individual level trial data can be used to capture heterogeneity of treatment effect. Multivariate outcome prediction models can be developed and validated using a wide range of techniques, including conventional regression models or machine-learning techniques such as causal forest models.[18]

Overall, we demonstrated that, in the presence of heterogeneity of treatment effect, differences in patient populations can explain a substantial portion of observed differences in outcomes across studies. Our proposed method, which leverages individual-level RCT data and uses an individual-level simulation model, can adjust for differences between populations and provide counterfactual outcomes of an RCT in a target population. Application of this method can facilitate interpretation of between-study variations in results and can support deriving better-informed inferences about true underlying treatment effects.

METHODS

Selected case studies

We selected as a case study the Randomized Evaluation of Long-Term Anticoagulation Therapy (RE-LY) trial, a large RCT that compared the efficacy and safety of dabigatran and warfarin in patients with atrial fibrillation.[19] We compared the RE-LY results with the results of two observational studies (Graham 2014 and Larsen 2013) that aimed at assessing the effectiveness and safety of dabigatran vs. warfarin for treatment of atrial fibrillation in routine care.[20,21] We chose this case study because HTE was observed in the RE-LY trial, a model examining modifiers of the treatment effect was available, and because several observational studies have examined similar outcomes.

Table 1 summarizes the baseline characteristics of patients in three studies. The main observed results of RE-LY trial and those of Graham 2014 and Larsen 2013 are summarized in **Table 2**. The median duration of the follow-up period was 2 years in RE-LY; dabigatran-treated and warfarin-treated patients contributed 12,042 and 11,796 person-years to the study, respectively. In the Graham 2014 study, dabigatran-treated and warfarin-treated patients contributed 18,205 and 19,382 person-years, respectively. The median follow-up time in Larsen 2013 was 10.5 months and the study included 4,086 person-years of dabigatran-treated patients.

All incidence rates in Graham 2014 are higher than those from the RE-LY trial, although the hazard ratios (HRs) were similar and their 95% confidence intervals (CIs) largely overlapped. For example, rates of ischemic stroke in Graham 2014 were 1.13 and 1.39 per 100 person-years for patients on dabigatran and warfarin (HR, 0.80; 95% CI, 0.67–0.96), respectively, compared to 0.92 and 1.20 per 100 person-years (HR, 0.76; 95% CI, 0.60–0.98) in RE-LY.

In contrast, the estimated incidence rates and HRs from Larsen 2013 differed substantially from the corresponding rates and HRs from the RE-LY trial. For example, incidence rates of ischemic stroke were 3.5 and 3.0 per 100 person-years (HR, 1.18; 95% CI, 0.85–1.64) for patients treated with dabigatran and warfarin, respectively, compared to 0.92 and 1.20 per 100 person-years (HR, 0.76; 95% CI, 0.60–0.98) in RE-LY.

Simulation model

We first developed a discrete event simulation model that used published outcomes rates from the RE-LY trial[22] as well as data on patients' baseline characteristics from RE-LY[19] to replicate the RCT results (**Figure S1**). Building the simulation model involved three main steps: 1) using a Monte Carlo simulation to generate hypothetical cohorts of patients with baseline covariate distributions that matched the marginal distribution of covariates in the RE-LY population; 2) designing a model structure based on *a priori* knowledge about possible pathways and health states that patients may experience over time; and 3) defining the relationship between patient-level baseline covariates (e.g., patient characteristics, event histories, treatment assignment) and the risk of outcomes based on published outcome risk models.

For the main analysis, the relations between patient level covariates and the outcomes of stroke/SE, ischemic stroke, ICH, and major bleeding were modeled using CHADS2-specific outcome rates reported by Oldgren *et al*.[22] Oldgren *et al*. published follow-on analysis of RE-LY that estimated rates of stroke and thromboembolism, intracranial hemorrhage, and major bleeding based on trial participants' CHADS2 scores.[22] The CHADS2 score is a simple, validated measure of risk that assigns two points for a history of stroke or transient ischemic attack, and one point for having a history of congestive heart failure, hypertension, age 75 years or older, and diabetes. Using RE-LY trial data, Oldgren *et al*. demonstrated that risk of thromboembolic and bleeding events varies by patients' CHADS2 score (**Table S2**). The results suggested increasing rates of thrombotic and bleeding events and corresponding HRs for higher categories of CHADS2 score. We used these results to predict risk of stroke and thromboembolism, intracranial bleeding, and major bleeding, given patients' characteristics in our simulation model. Possible health states and pathways in the model were defined based on *a priori* clinical knowledge about possible outcomes of patients with atrial fibrillation treated using anticoagulants.

Comparison of predicted RCT results to observed RCT results

We simulated hypothetical cohorts of 6,000 warfarin patients and 6,000 dabigatran patients with the marginal covariate distributions and sample size similar to those of the patients in the RE-LY trial (**Table S3**). The simulation model estimated predicted outcomes of RCT patients $\hat{O}_i$

RCT conditional on their characteristics C_i^{RCT}, assuming that the outcome prediction models and model structure accurately reflected the true relationship between covariates and outcomes: $\hat{O}_i RCT = f(C_i^{RCT})$, where $\hat{O}_i RCT$ is the expected outcome of interest, C_i^{RCT} is a vector of covariates for patient i in the RCT, and f represents all model assumptions that defined the relationship between patient level covariates and outcomes.

To validate the discrete event simulation model, we compared the expected outcomes $\hat{O}_i RCT$ with the observed RCT outcomes (O^{RCT}). If the model structure, f, is specified correctly, the model will accurately replicate the overall RCT outcomes (O^{RCT}) for a cohort of patients similar to RCT participants. We defined *modeling error* as the difference between observed and predicted results of RCT (i.e., $O^{RCT} - \hat{O}^{RCT}$ for each particular outcome (**Table S4**)).

Comparison of predicted observational results to observed RCT results

Next, we simulated predicted outcomes for patients in the observational studies. We used the discrete event simulation model developed and validated among RCT patients to estimate expected outcomes in the observational study populations: $O^{obs} = f(C^{obs})$. This was done by changing the baseline characteristics of the simulated cohort to match marginal distribution of the patients' characteristics in each of the observational studies. For each of those observational studies, we simulated hypothetical cohorts of 6,000 warfarin patients and 6,000 dabigatran patients. As such, the simulated outcomes $\hat{O}^{obs}$ can be interpreted as the results of the RCT had it been conducted in a population similar to that in the observational study. *HTE* was defined to be equal to the difference between predicted result of RCT and predicted result of observational study $[\hat{O}^{RCT} - \hat{O}^{obs}]$. This component reflects variation in results that are likely caused by the differences in study populations.

Comparison of predicted observational results to observed observational results

We also defined the difference between predicted and observed results of observational studies $[\hat{O}^{obs} - O^{obs}]$ as *unexplained discrepancies*. Based on this definition, all discrepancies between observed results of RCT and observed results of observational studies that could not be explained by modeling error or HTE were defined as unexplained discrepancies. Therefore, unexplained discrepancies include factors such as confounding, misclassification of outcomes, differences in outcomes and risk factors measurement, differences in follow-up, and differences in adherence in RCT vs. electronic databases. Note that, based on this definition, modeling error, HTE, and unexplained components add up to be equal to differences between observed results of RCT an observational studies. These notations and concepts are summarized in **Table S4**.

Patient involvement

No patients were involved in this study. Defining the research question and the outcomes, study design, implementation, interpretation, and reporting of the results were done solely by study investigators.

FUNDING

No funding was received for this work.

Additional Supporting Information may be found in the online version of this article.

CONFLICT OF INTEREST

The authors declare no competing interests for this work. Dr. Najafzadeh is the Principal Investigator of an investigator-initiated unrestricted grant from Baxalta to Brigham and Women's Hospital, unrelated to the topic of this study. He also was a consultant to AltheaDx for a project unrelated to this study. Dr. Gagne is the Principal Investigator of an investigator-initiated unrestricted grant from Novartis Pharmaceuticals to Brigham and Women's Hospital for unrelated work. He is a consultant to Aetion, a software company, and to Optum. Dr. Schneeweiss is a consultant to WHISCON and to Aetion, a software manufacturer of which he also owns shares. He is principal investigator of investigator-initiated grants to the Brigham and Women's Hospital from Novartis, and Boehringer Ingelheim unrelated to the topic of this study. Dr. Wang is a consultant to Aetion and is supported by grant number R00HS022193 from the Agency for Healthcare Research and Quality, unrelated to this study.

AUTHOR CONTRIBUTIONS

M.N., S.S., N.K.C., S.W., and J.G. wrote the article; M.N. and J.G. designed the research; M.N. and J.G. performed the research; M.N. analyzed the data; M.N., J.G., S.S., S.V., and N.K.C. contributed new reagents/analytical tools.

REPRODUCIBLE RESEARCH STATEMENT

Study protocol: Not applicable. Statistical code: Available from Dr. Najafzadeh (e-mail, mnajafzadeh@bwh.harvard.org). Dataset: Input parameters and sources are provided in the text.

1. Concato, J., Shah, N. & Horwitz, R.I. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* **342**, 1887–1892 (2000).
2. Dahabreh, I.J. & Kent, D.M. Can the learning health care system be educated with observational data? *JAMA* **312**, 129–130 (2014).
3. Ioannidis, J.P. *et al*. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* **286**, 821–830 (2001).
4. Schneeweiss, S. *et al*. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Med. Care* **45**, S131 (2007).
5. Rothwell, P.M. External validity of randomised controlled trials:"To whom do the results of this trial apply?" *Lancet* **365**, 82–93 (2005).
6. Schneeweiss, S., Huybrechts, K. & Gagne, J. Interpreting the quality of health care database studies on the comparative effectiveness of oral anticoagulants in routine care. *Comp. Eff. Res.* **3**, 33–41 (2013).
7. Cole, S.R. & Stuart, E.A. Generalizing evidence from randomized clinical trials to target populations the actg 320 trial. *Am. J. Epidemiol.* **172**, 107–115 (2010).
8. Stuart, E.A., Bradshaw, C.P. & Leaf, P.J. Assessing the generalizability of randomized trial results to target populations. *Prevent. Sci.* **16**, 475–485 (2015).
9. Basu, S., Hong, A. & Siddiqi, A. Using decomposition analysis to identify modifiable racial disparities in the distribution of blood pressure in the United States. *Am. J. Epidemiol.* **182**, 345–353 (2015).
10. Oaxaca, R. Male-female wage differentials in urban labor markets. *Int. Econ. Rev.* 693–709 (1973).
11. Reilly, P.A. *et al*. The effect of dabigatran plasma concentrations and patient characteristics on the frequency of ischemic stroke and major bleeding in atrial fibrillation patients: the RE-LY trial (randomized evaluation of long-term anticoagulation therapy). *J. Am. Coll. Cardiol.* **63**, 321–328 (2014).
12. Najafzadeh, M. & Schneeweiss, S. From trial to target populations — calibrating real-world data. *N. Engl. J. Med.* **376**, 1203–1205 (2017).
13. Bierer BE, Li R, Barnes M, Sim I. A global, neutral platform for sharing trial data. *N. Engl. J. Med.* **374**, 2411–2413 (2016).
14. Taichman, D.B. *et al*. Sharing clinical trial data: a proposal from the international committee of medical journal editors. *Lancet* **387**, e9–11 (2016).
15. Dahabreh, I.J., Hayward, R. & Kent, D.M. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *Int. J. Epidemiol.* **45**, 2184–2193 (2016).
16. Burke, J.F., Hayward, R.A., Nelson, J.P. & Kent, D.M. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. *Circ. Cardiovasc. Qual. Outcomes* **7**, 163–169 (2014).
17. Kent, D.M. *et al*. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. *Int. J. Epidemiol*. **45**, 2074–2088 (2016).

18. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 7353–7360 (2016).
19. Connolly, S.J. *et al*. Dabigatran versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* **361**, 1139–1151 (2009).
20. Graham, D.J. *et al*. Cardiovascular, bleeding, and mortality risks in elderly medicare patients treated with dabigatran or warfarin for nonvalvular atrial fibrillation. *Circulation* **131**, 157–164 (2015).
21. Larsen, T.B. *et al*. Efficacy and safety of dabigatran etexilate and warfarin in "real-world" patients with atrial fibrillation: a prospective nationwide cohort study. *J. Am. Coll. Cardiol.* **61**, 2264–2273 (2013).
22. Oldgren, J. *et al*. Risks for stroke, bleeding, and death in patients with atrial fibrillation receiving dabigatran or warfarin in relation to the chads2 score: a subgroup analysis of the RE-LY trial. *Ann. Intern. Med*. **155**, 660–667 (2011).