PASS information

Title	Estimating Oral Anticoagulant Comparative Effectiveness in the Setting of Effect Heterogeneity: Comparing Clinical Trial Transport and Observational Epidemiologic Methods
Protocol version	V1.0
Date of last version of protocol	8/31/2018
EU PAS register number	
Active substance	Dabigatran etexilate and vitamin K antagonists
Medicinal product	Pradaxa
Product reference	N/A
Procedure number	N/A
Marketing authorization holder	While Boehringer Ingelheim owns the patent for Pradaxa, this study is being conducted independently.
Joint PASS	No
Research question and objective	This study is designed to estimate the absolute effects of dabigatran use relative to warfarin on death, stroke, and bleeding in a cohort of older

	adults in routine care by 1) using weights to standardize the trial population to the older adults and 2) use more traditional observational epidemiologic methods in the population of older adults. These results will then be contrasted with one another, and the data will additionally be used to explore potential problems of misclassification and missing data when using weights to standardize trial populations.
Country of study	United States
Author	Michael Webster-Clark Email: mawc@live.unc.edu
Marketing authorization holder	N/A
MAH contact person	N/A

1. Table of contents:

1. Table of contents……p.3

2. List of abbreviations……p.5

3. Responsible parties……p.7

4. Abstract……p.8

5. Amendments and updates……p.10

6. Milestones……p.11

7. Rationale and background……p.12

8. Research question and objectives……p.18

9. Research methods……p.20

9.1. Study design……p.20

9.2. Setting……p.21

9.3. Variables……p.24

9.4. Data sources……p.33

9.5. Study size……p.39

9.6. Data management……p.40

9.7. Data analysis……p.41

9.8. Quality control……p.52

9.9. Limitations of the research methods……p.52

9.10. Other aspects……p.53

10. Protection of human subjects……p.54

11. Management and reporting of adverse events/adverse reactions…….p.55

12. Plans for disseminating and communicating study results……p.56

13. References……p.57

14. **Annex 1. List of stand-alone documents……p.60**
15. **Annex 2. ENCePP checklist for study protocols……p.61**
16. **Annex 3. Additional information……p.62**

2. List of abbreviations

AF = atrial fibrillation

$CHADS_2$ = stroke risk score based upon Congestive heart failure, Hypertension, Age, Diabetes, and past stroke

CHA_2DS_2Vasc = stroke risk score from $CHADS_2$ and vascular disease

C.I. = confidence interval

CPT = current procedural technology

CSDR = Clinical Study Data Request

CTDT = Clinical Trial Data Transparency (platform)

EMM = effect measure modification or effect measure modifier

HR = hazard ratio

INR = international normalized ratio

MCAR = missing completely at random

MAR = missing at random

MNAR = missing not at random

MiscCAR = misclassification completely at random

MiscAR = misclassified at random

MiscNAR = misclassified not at random

NOAC = novel oral anticoagulant

RCT = randomized controlled trial

RD = risk difference

RR = risk ratio

sIOSW = stabilized inverse odds of sampling weights

sIPCW = stabilized inverse probability of censoring weights

sIPTW = stabilized inverse probability of treatment weights

3. Responsible parties

Principal investigator: Michael Webster-Clark, PharmD. Contact: mawc@live.unc.edu

4. Abstract

Title: Estimating Oral Anticoagulant Comparative Effectiveness in the Setting of Effect Heterogeneity: Comparing Clinical Trial Transport and Observational Epidemiologic Methods. Michael Webster-Clark, University of North Carolina at Chapel Hill, v.1.0

Rationale and background: Atrial fibrillation affects 33 million adults worldwide.[1] Even if individuals with atrial fibrillation are asymptomatic, stroke incidence in the atrial fibrillation population is much higher and resulting strokes are more frequently associated with death, hospitalization, and long-term disability than strokes in adults without atrial fibrillation.[2, 3] Warfarin, the historical standard treatment for preventing strokes in atrial fibrillation, is difficult to manage therapeutically due to its lengthy half-life and narrow therapeutic range. Warfarin overdose can also result in catastrophic bleeding events.[4] Novel oral anticoagulants are easy to manage with simple dose adjustments for renal insufficiency and have been shown to be non-inferior to warfarin administered with systematic management protocols in clinical trial populations.[5] One of the first novel anticoagulants to be approved in the United States, dabigatran, was shown to be more effective than warfarin at stroke and embolic event prevention (HR 0.66, 95% C.I. 0.53-0.82) with no increase in bleeding (HR 0.93, 95% C.I. 1.07) in the RE-LY trial.[6]

Research questions: Estimates of efficacy in these clinical trials are likely not perfect estimates of effectiveness in clinical care.[7, 8] Patients selected into trials tend to be young with fewer comorbidities than the general population; this can modify the population average treatment effect.[9, 10] To address concerns about this potential treatment effect modification, studies have used observational claims data to directly estimate safety of novel oral anticoagulants compared to warfarin in clinical care and observed attenuated efficacy and differing safety profiles. Unfortunately, their results may be confounded by unmeasured variables.[11-20] Furthermore, warfarin management protocols from trials do not necessarily represent the way warfarin is managed for patients in routine clinical care, making it difficult to know how consistent warfarin treatment is between trial and observational populations.[21] Finally, it is

unclear how much misclassified and missing data in key effect modifiers (a common issue in claims data)[22] could bias estimates when transporting results to new target populations.

Study design: We will be conducting two main studies in this analysis with two different study designs: a retrospective observational study embedded in the United States' 20% Medicare sample, and a re-analysis of the RE-LY trial data after weighting the trial to resemble the target population of Medicare beneficiaries while maintaining randomization.

Populations and Data Sources: This study will use two data sources from two populations that are collected in very different ways: the RE-LY trial's individual-level data and cohorts of initiators of dabigatran and warfarin in Medicare from 2010 to October 2015.
Variables: This study will include a wide variety of potential modifiers of treatment effect in the weighting analysis, including use of potentially interacting medications, comorbidities, sex, and age. These variables and more will be included in the analyses entirely within the observational data.

Study size: We anticipate a cohort of approximately 157,000 Medicare beneficiaries based upon extrapolation from the FDAs matched study in the 100% Medicare sample, with approximately 12,000 individuals from RE-LY contributing data.[15]

Data analysis: To address concerns about potential treatment effect attenuation, we will estimate the absolute scale causal effect of dabigatran compared to warfarin using inverse odds of sampling weights (IOSW) to transport effect estimates from RE-LY and use outcome data on the Medicare atrial fibrillation cohort to check whether consistency and exchangeability assumptions hold.[23-26] We will separately generate an effect estimate in the routine Medicare atrial fibrillation cohort using propensity score weighting using only the observational data for comparison with the transported estimate. Finally, we will manipulate the observed RE-LY and Medicare populations to assess when misclassified and missing data lead to bias or imprecision of transported treatment effect estimates in real-world data.

Milestones: We aim to have a paper documenting the results from the observational cohort submitted to a peer-reviewed journal by January 2018, with another paper comparing the weighted RE-LY trial results by June 2018.

5. Amendments and updates

Number	Date	Section of study protocol	Amendment or update	Reason
1.0	8/31/2018	All	Created	Submission prep

6. Milestones

Milestone	Date
Data access obtained (Medicare)	9/2018
Data access obtained (RE-LY)	10/2018
Observational study results	1/2019
Weighted trial study results	4/2019
Misclassification and missing data study results	6/2019

7. Rationale and background

Atrial fibrillation affects 33 million adults worldwide.[1] Even if individuals with atrial fibrillation are asymptomatic, stroke incidence in the atrial fibrillation population is much higher and resulting strokes are more frequently associated with death, hospitalization, and long-term disability than strokes in adults without atrial fibrillation.[2, 3] Warfarin, the historical standard treatment for preventing strokes in atrial fibrillation, is difficult to manage therapeutically due to its lengthy half-life and narrow therapeutic range. Warfarin overdose can also result in catastrophic bleeding events.[4] Novel oral anticoagulants are easy to manage with simple dose adjustments for renal insufficiency and have been shown to be non-inferior to warfarin administered with systematic management protocols in clinical trial populations.[5] One of the first novel anticoagulants to be approved in the United States, dabigatran, was shown to be more effective than warfarin at stroke and embolic event prevention (HR 0.66, 95% C.I. 0.53-0.82) with no increase in bleeding (HR 0.93, 95% C.I. 1.07) in the RE-LY trial.[6]

LITERATURE REVIEW

We conducted a literature review of the randomized controlled trials and observational studies comparing dabigatran to warfarin, starting with randomized controlled trials.

Randomized controlled trials:

Dabigatran's approval for the indication of stroke prophylaxis in AF was based on the RE-LY non-inferiority trial.[6] In RE-LY, more than 18,000 AF patients from 951 clinical centers in more than 44 countries were randomized to receive warfarin under a standard dosing protocol, twice daily dabigatran 110 mg, or twice daily dabigatran 150 mg. Patients were blinded to which dose of dabigatran they received, but use of warfarin was open-label. Patients were followed for a variety of outcomes in an intention-to-treat analysis. Markedly reduced hazards were observed for the primary efficacy outcome of stroke or systolic embolism for patients in the 150 mg dabigatran arm versus the warfarin arm (HR: 0.66, 95% C.I. 0.53, 0.82) and major bleeding overall was also slightly lower (HR: 0.93, 95% C.I. 0.81, 1.07), but much higher rates of gastrointestinal bleeding were observed (HR: 1.50, 95% C.I. 1.19, 1.89). Notably, the HR for

ischemic stroke (rather than combined hemorrhagic and ischemic stroke) was 0.76 (95% C.I. 0.60, 98) in the 150 mg arm. The 110 mg dose of dabigatran, on the other hand, had a larger decrease in major bleeding versus warfarin (HR: 0.80, 95% C.I. 0.69, 0.93) and a smaller increase in gastrointestinal bleeding (HR: 1.10, 95% C.I. 0.86, 1.41) but a much smaller improvement in the primary efficacy outcome (HR: 0.91, 95% C.I. 0.74, 1.11). Both the 110 mg and 150 mg dosage improved survival versus warfarin with a HR of 0.91 (95% C.I. 0.80, 1.03) and 0.88 (95% C.I. 0.77, 1.00), respectively. The non-inferiority margins set by the investigators were all met, and the drug was approved for dosing at 150 mg twice daily. It is not clear, however, how these tradeoffs might manifest in populations with differing distributions of risk factors for bleeding and stroke.

Observational studies:

There have been several comparative effectiveness studies published using observational and claims data in an attempt to determine whether dabigatran and other NOACs are as or more effective in practice compared to their performance in clinical trials. Their results have varied, but all have differed from RE-LY in one way or another. On the whole they reinforce the need for additional studies and examination of potential treatment effect heterogeneity.

Perhaps the largest study to date has been the FDA's analysis using initiators of warfarin and dabigatran in the full Medicare sample from October 2010 to December 2012.[15] They estimated a hazard ratio of 0.80 (95% C.I. 0.67, 0.96) for ischemic stroke, favoring dabigatran, and a hazard ratio of 0.97 (95% C.I. 0.88, 1.07) for major bleeding, with a heightened risk of gastrointestinal bleeding for dabigatran patients with a hazard ratio of 1.28 (95% C.I. 1.14, 1.44); when they looked at dosages of 150 mg specifically, rather than 75 mg, they found improved stroke reduction (hazard ratio of 0.70) but greater risk of gastrointestinal bleed (hazard ratio of 1.51) more in line with the results of the RE-LY trial. They did notice, however, that the increase in the risk of gastrointestinal bleeding was concentrated in older adults, particularly women over 75 and men over 85. Women over 85 were the only group with an increased risk of mortality from dabigatran, but other groups experienced an improvement in

mortality with an overall hazard ratio of 0.76 (95% C.I. 0.67, 0.86) with the 150 mg dose, more pronounced than the 0.88 observed in the RE-LY trial.

At the same time the FDA study was running on the full Medicare sample, Hernandez et al were investigating the question using the 5% Medicare sample.[16] This study only used Medicare initiators from October 2010 to October 2011, limiting the population, and required two outpatient diagnoses for AF or atrial flutter to qualify an individual as an AF patient, as well as requiring individuals to fill within two months of their incident diagnosis (while the FDA merely required a diagnosis at any time before filling). They focused predominantly on bleeding outcomes, rather than stroke or systemic embolism.. After adjustment, they identified an increased risk of major bleeding with a hazard ratio of 1.58 (95% C.I. 1.36, 1.83) and an increased risk of gastrointestinal bleeding with a hazard ratio of 1.85 (95% C.I. 1.64, 2.07), both higher than the original RE-LY trial. Despite being conducted on a subsample of the FDA data, their results looked quite a bit worse for dabigatran, suggesting that either the difference in grace period, the use of a different target population than the FDA (the whole population of initiators rather than the dabigatran patients targeted in a matched design), worse confounding control, or all three.

Other U.S. governmental databases have been used to answer this question as well. Villines et al utilized the Department of Defense database, which provides uniform medical coverage and pharmacy benefits to nearly 10 million individuals receiving care at both military and non-military institutions, to find initiators of dabigatran and warfarin from October 2010 to July 2012 with at least one diagnosis for AF within 12 months of their initiation.[17] They identified a cohort of 14,813 dabigatran users and 24,500 warfarin initiators, with propensity score matching using a model built by backwards selection reducing it to 12,793 of each. Comparing dabigatran to warfarin, they estimated a hazard ratio for stroke of 0.73 (95% C.I. 0.55, 0.97), a hazard ratio for major bleeding of 0.87 (95% C.I. 0.74, 1.03) and 0.82 (95% C.I. 0.71, 0.95) when restricting to 150 mg doses, and a hazard ratio for gastrointestinal bleeding of 1.13 (95% C.I. 0.94, 1.37). Additionally, they found a hazard ratio for myocardial infarction of 0.65 (95% C.I. 0.45, 0.95), much lower than that observed in RE-LY of 1.35 (95% C.I. 0.98, 1.87); similarly, the hazard ratio for death was 0.64 (95% C.I. 0.55, 0.74).

Other researchers examined private insurance databases. Seeger, Schneeweiss, et al performed an analysis using data from two commercial insurance databases (MarketScan from Truven and Clinformatics from Optum) focusing on both safety and efficacy.[18] These investigators identified 41,103 warfarin and 18,560 dabigatran initiators between October 2010 and December 2013. The authors matched on a propensity score with 78 investigator-specified covariates, resulting in a final cohort of 15,529 initiators of each medication (with successful removal of imbalances in the covariates as measured by the SMD), with an additional analysis using high-dimensional propensity scores. This study was the first to look at benefits on an absolute scale, rather than reporting purely hazard ratios. Individuals stopped follow-up at the time of death. When contrasting dabigatran with warfarin, they found a hazard ratio of 0.77 (95% C.I. 0.54, 1.09) and one-year risk difference of -0.0003 (95% C.I. -0.0006, 0.0002) for strokes and a hazard ratio of 0.75 (95% C.I. 0.65, 0.87) and risk difference of -0.018 (95% C.I. -0.025, -0.010) for major hemorrhages; they did not investigate gastrointestinal hemorrhages or mortality.

Another large-scale study was conducted in Denmark, where both the 110 mg and 150 mg twice daily doses of dabigatran were approved for use in August 2011.[27] They included dabigatran initiators after August 2011 and only allowed warfarin initiators to enter between August 2009 and July 2010, resulting in initial cohorts of 5,106 dabigatran and 13,548 warfarin patients. This was another propensity-score matched analysis (this time matched 2:1) that found matches for 4,978 dabigatran patients of either dosage and censored at treatment switching but not discontinuation with two levels of propensity score. They found large mortality benefits for both doses of dabigatran with hazard ratios of 0.79 (95% C.I. 0.65, 0.95) for the 110 mg dose and 0.57 (95% C.I. 0.40, 0.80) for the 150 mg dose compared to warfarin. They also found much lower risks of gastrointestinal bleeding comparing dabigatran 110 mg with warfarin with a hazard ratio of 0.60 (95% C.I. 0.37, 0.93) and only a slight increase for 150 mg with a hazard ratio of 1.12 (95% C.I. 0.67, 1.83), with favorable results for major bleeding with both dosages as well (HR: 0.82 (95% C.I. 0.59, 1.12) for the 110 mg dose and HR: 0.77 (95% C.I. 0.51, 1.13) for the 150 mg dose). Stroke benefits were also inconsistent with findings from other studies and with the logical assumption that higher doses will prevent more embolisms,

with hazard ratios of 0.73 (95% C.I. 0.53, 1.00) for the 110 mg dose and 1.18 (95% C.I. 0.85, 1.64) for the 150 mg dose. Given the large differences between their cohorts even after propensity score matching and the strange trends between dabigatran doses (which could be the result of matching to different cohorts), their results are difficult to interpret.

The final and most recent study focusing specifically on the contrast between dabigatran and warfarin in AF was conducted by the FDA with the Sentinel system in 2017.[20] The Sentinel network collects data from a variety of administrative, clinical, and pharmacy dispensing databases for use in large-scale investigations of key medical questions in the United States.[28] Using this system and data from November 2010 to May 2014, Go et al conducted an propensity-score matched analysis (with matching and model estimation performed within each of the data partners), identifying 25,289 dabigatran initiators and finding matches for each one amongst the 83,034 warfarin initiators. They estimated a hazard ratio for ischemic stroke of 0.92 (95% C.I. 0.65, 1.28) and a hazard ratio for gastrointestinal bleeding of 1.04 (95% C.I. 0.83-1.30), meaning their results showed both less benefit and less harm than the RE-LY trial. Interestingly, however, they did identify substantial heterogeneity in gastrointestinal bleeding with those under 65 having a hazard ratio of 0.59 (95% C.I. 0.32-1.07), those between 65 and 74 having a hazard ratio of 0.81 (95% C.I. 0.52-1.24), those between 75 and 84 having a hazard ratio of 1.47 (95% C.I. 1.05, 2.14) and those over 85 experiencing a hazard ratio of 1.84 (95% C.I. 1.05, 3.20). There was also some heterogeneity by kidney function in both ischemic stroke and gastrointestinal bleeding risks, though this may be due to other concomitant factors associated with reduced kidney function like age, hypertension, diabetes, and congestive heart failure.

There is another study focusing on several different NOACs with dabigatran as one of the potential options whose results also warrant discussion. Lip et al conducted a study in MarketScan examining NOAC initiators from January to December 2013 focusing specifically on major bleeding risk and using Cox proportional hazards with direct adjustment for a variety of variables and backwards selection at $p < 0.2$.[11] These authors identified a decreased rate of major bleeding for dabigatran (HR: 0.88, 95% C.I. 0.64-1.21) for dabigatran relative to warfarin. Their propensity matched analysis showed a slightly reduced rate of major bleeding with a hazard ratio of 0.69 (95% C.I. 0.50-0.96), suggesting some potential treatment effect

heterogeneity.[12] In both analyses individuals were followed until discontinuation from their initial medication or switching, though the amount of gap or grace period they allowed is unclear. These results generally seem to agree with those of other studies in younger, claims-based cohorts.

Overall, there is still significant clinical uncertainty about the actual benefit/risk profile of dabigatran compared to warfarin; still, these studies have clearly shed light on the fact that there is likely heterogeneity in the safety and efficacy of dabigatran with respect to warfarin. While each of the studies added to our body of knowledge only one of them implemented an intention-to-treat analysis for comparison with RE-LY. Beyond this, almost all of the studies focused on estimating a treatment effect amongst dabigatran initiators, who often have characteristics that set them apart from the general patients included in RE-LY. Additionally, most stopped following individuals at their time of death while using Cox proportional hazards and Kaplan-Meier estimators that effectively assume we can prevent competing risks; this can be problematic, particularly when the competing event is roughly as common or more common than the event of interest.[29] A new study in a Medicare cohort spanning more time using both IPTW and matched designs alongside an attempt to create an analysis using the trial data that conditions on staying on therapy (while accounting for differences in potential effect measure modifiers among individuals switching therapies) would significantly improve understanding of which medication performs better in an older population with a higher prevalence of comorbid conditions, without having to assume that adherence patterns will be constant across the RE-LY and Medicare cohorts with respect to baseline modifiers.

8. Research question and objectives

The purpose of this project is to estimate the effect of dabigatran versus initiation on the risk of stroke, bleeding, and mortality in a real-world target population using trial transport and observational methods and further explore the impact of missing data and misclassification on said trial transport methods.

We will:

Specific Aim 1: Estimate the effect of dabigatran versus warfarin initiation on the 1 and 2-year risk of stroke, bleeding, and death among Medicare beneficiaries with atrial fibrillation using transportability methods reweighting the RE-LY trial data.

> **1.1:** Compare risks of each outcome between the RE-LY and Medicare cohorts on each treatment before and after using inverse odds of sampling weights to standardize RE-LY to Medicare with respect to effect measure modifiers (EMM).
>
> **1.2:** Estimate the absolute effect of dabigatran versus warfarin on 1 and 2-year risks of stroke, bleeding, and death using inverse odds of sampling weights to standardize the RE-LY trial to the Medicare cohort in effect measure modifiers.

Specific Aim 2: Estimate the effect of dabigatran versus warfarin initiation on 1 and 2-year risks of stroke, bleeding, and death in Medicare beneficiaries with atrial fibrillation using observational data and propensity score methods for comparison with the results from Aim 1.

> **2.1:** Estimate treatment effects in warfarin and dabigatran new users with a new user active comparator design.
>
> **2.2:** Estimate treatment effects in dabigatran new users and dabigatran switchers with a prevalent user design.
>
> **2.3:** Compare these estimates and their precision to those obtained using the clinical trial data.

Specific Aim 3: Explore the effects of creating missing data and misclassification on the target population effect estimates of the 1 and 2-year risk differences for stroke, bleeding, and death and their precision using covariate data simulated from the RE-LY trial and Medicare cohort data.

3.1: Assess the effects of different methods of addressing missing data on the transported effect estimates and their precision after simulating various types of covariate missingness in the RE-LY and Medicare data.

3.2: Assess the effect of ignoring or addressing misclassification on the transported effect estimates and their precision after simulating various types of covariate misclassification in the RE-LY and Medicare data.

9. Research methods

This research plan is designed to compare the risk of stroke and bleeding in older adults in routine care taking warfarin or dabigatran with risks in RE-LY trial participants, before and after adjusting for known risk factors; estimate the effects of dabigatran compared to warfarin on stroke and bleeding risk in older adults on the absolute scale using transport weights; assess the degree to which estimates from various observational methods agree with these transported trial results; and evaluate how missing or misclassified effect measure modifiers in trials and claims data alter transported estimates.

9.1 Study design

The aims call for estimating the safety and effectiveness of initiating and staying on dabigatran compared to starting and staying on warfarin in the U.S. Medicare population with atrial fibrillation (AF) through transporting the results of the RE-LY trial to several target populations defined by treatment initiation. We will be analyzing effects on absolute risks of stroke, bleed, and death at 1- and 2-years after dabigatran or warfarin initiation, if possible. Because standard observational approaches and direct use of the trial estimate both have their limitations, we will need to fuse cohort data from the Medicare population with the RE-LY trial data. Combining these two study designs makes it possible to estimate average treatment effects in the Medicare population. This estimate will be valid under the assumption that a sufficient set of causes of sampling into the trial that are risk factors for the outcome have been accounted for.[30] The outcome data collected on patients in Medicare will allow us check this assumption using various weighting techniques as well as assess how different the results are from results obtained using observational approaches like the new user active comparator or prevalent user design. Finally, having access to both cohorts will allow us to explore the change in estimates and their precision resulting from misclassification and missing data by censoring and misclassifying real observed data rather than relying on pure simulation approaches.

9.2 Setting

This study will use two data sources from two populations that are collected in very different ways: the RE-LY trial's individual-level data and cohorts of initiators of dabigatran and warfarin in Medicare from 2010 to October 2015.

RE-LY Population Description

Individual-level data for the RE-LY trial will be obtained using Clinical Study Data Request (CSDR) and analyzed on the Clinical Trial Data Transparency (CTDT) platform, soliciting data access as soon as the analysis protocol is finalized. The RE-LY trial randomized 18,113 patients to warfarin, 110 mg dabigatran twice daily, and 150 mg of dabigatran twice daily. We will focus primarily on the results for the 150 mg dosage of dabigatran, since the 110 mg dose was not approved for usage in the United States, though we could also integrate the 110 mg dose in secondary analyses. Eligible patients had documented AF during the six months before their enrollment, as well as at least one of five other risk factors for stroke including: age greater than 75; previous stroke, TIA, or systemic embolism; left ventricular ejection fraction under 40% in the past six months; a diagnosis of diabetes mellitus and age over 65; or hypertension requiring pharmaceutical treatment and age over 65; or documented coronary artery disease and age over 65.

Key exclusions to assure either safety or efficacy of warfarin or dabigatran included reversible AF, prosthetic heart valves or other conditions for which dabigatran had not been tested, stroke within the past 14 days or severe stroke within the past 6 months, a variety of conditions associated with increased risk of bleeding, active infective endocarditis, active liver disease, anemia or thrombocytopenia, patients judged unreliable or having a life expectancy less than the expected trial duration, patients who received another investigational drug within 30 days, transaminase elevations in response to ximelagatran (another agent with a similar mechanism of action), and patients with severe renal impairment (creatinine clearance of equal to or less than 30 mL/min). The RE-LY trial was conducted in a population from a variety of countries with a wide range of ages (mean age in dabigatran 150 mg of 71.5 years, standard deviation 8.8) and

collected data on multiple potential causes of heterogeneity including medication use, past stroke, and other medical diagnoses.

As a randomized controlled trial there is no confounding in expectation in the baseline covariate distribution in RE-LY (and analyses suggested limited chance confounding). This allows estimation of an internally valid intention-to-treat effect estimate in the target population using outcome data from RE-LY in Aim 1 provided we have weighted the RE-LY data to match the distributions of measured effect modifiers in our target. It does not, however, guarantee that an estimate censoring at treatment discontinuation will be internally valid. We will limit our population to patients over 65 from the RE-LY trial to ensure we are looking only at older adults; this is not problematic from a sample size perspective, as 85% of the initial RE-LY trial population was over the age of 65.

Medicare Atrial Fibrillation Population Description

The specific Medicare data used will be the 20% sample of all Medicare beneficiaries with fee-for-service coverage of Medicare Parts A, B, and D for at least one month from 2007-2015, available at the University of North Carolina through the Sheps Center. We will be constructing two main study cohorts from the Medicare population from 2010 (when dabigatran was approved for use in the United States) to October 2015 (when the United States transitioned from ICD-9 to ICD-10 codes) for this analysis: first, a cohort of dabigatran and warfarin initiators; and second, a cohort of dabigatran and warfarin initiators, where dabigatran users are allowed to have previously initiated warfarin or switched to dabigatran.

Patients will be eligible for inclusion into the Medicare cohort at their first initiation of warfarin or 150 mg twice daily dabigatran with a 2-month washout period for use of either drug or another NOAC (apixaban, rivaroxaban, and edoxaban), provided they have an inpatient or outpatient AF diagnosis code in the 6 months before or 1 week after their prescription (in which case follow-up will begin at the time of diagnosis). This requirement for a recent AF diagnosis code is analogous to the RE-LY trial inclusion criteria that required evidence of recent AF. Individuals will also have to fit the eligibility criteria for RE-LY including at least one risk factor for stroke (described in the **RE-LY Population Description**), as represented by at least one

diagnosis code in the year prior to initiation. Individuals will also need to have 12 months continuous coverage in Medicare parts A, B, and D before this index prescription to enable assessment of eligibility criteria, key effect measure modifiers, and potential confounders. This period will also be used to exclude individuals with identifiable exclusion criteria for the trial, including liver disease, severe stroke within the past six months, anemia and thrombocytopenia, valvular AF or prosthetic heart valves, and severe renal insufficiency (specific codes listed in **Appendix B**). If there are multiple initiations, only the first eligible initiation will be included. This Medicare cohort will contribute external validity to the project and allow insight into the distribution of effect modifiers in a general clinical cohort participating in anticoagulation care that may not be willing or able to participate in a full randomized trial, critical for **Aim 1**. It will also provide data for the observational analysis in **Aim 2** and the assessment of the effects of missing data in **Aim 3**.

The Medicare cohort will be further divided into four potential target populations: 1) all patients initiating warfarin for AF that could have been included in RE-LY; 2) all patients initiating dabigatran for AF that could have been included in RE-LY; 3) all patients initiating either drug for AF that could have been included in RE-LY; and 4) all dabigatran initiators and switchers that could have been included in RE-LY.

9.3 Variables

We will be using the definitions described below for each covariate in the Medicare cohort, but maintaining the adjudicated and clinically reviewed definitions from the RE-LY trial in those participants. Exposure will be defined based upon RedBook's NDC-generic name linkage.

Outcome	**Hospital Discharge Codes**
Stroke	As primary discharge diagnosis: 431.x Intracerebal hemorrhage 433.x1 Occlusion and stenosis of precerebral arteries 434.x1 Occlusion and stenosis of cerebral arteries with cerebral infarction 436.x Acute but ill-defined cerebrovascular events
Major bleeding	As primary discharge diagnosis: *Intracranial bleeding:* 430.x Subarachnoid hemorrhage 431.x Intracerebral hemorrhage 432.x Other and unspecified intracranial hemorrhage *Upper gastrointestinal bleed:* 531.0x Acute gastric ulcer with hemorrhage with or without obstruction 531.2x Acute gasric ulcer with hemorrhage and perforation with or without obstruction 531.4x Chronic or unspecified gastric ulcer with hemorrhage with or without obstruction 531.6x Gastric ulcer with hemorrhage and perforation with or without obstruction

	532.0x Acute duodenal ulcer with hemorrhage with or without obstruction 532.2x Acute duodenal ulcer with hemorrhage and perforation with or without obstruction 532.4x Chronic or unspecified duodenal ulcer with hemorrhage with or without obstruction 532.6x Chronic or unspecified duodenal ulcer with hemorrhage and perforation with or without obstruction 533.0x Acute peptic ulcer of unspecified site with hemorrhage with or without obstruction 533.2x Acute peptic ulcer of unspecified site with hemorrhage and perforation with or without obstruction 533.4x Chronic or unspecified peptic ulcer of unspecified site with hemorrhage with or without obstruction 533.6x Chronic or unspecified peptic ulcer of unspecified site with hemorrhage and perforation with or without obstruction 534.0x Acute gastrojejunal ulcer with hemorrhage with or without obstruction 534.2x Acute gastrojejunal ulcer with hemorrhage and perforation with or without obstruction 534.4x Chronic or unspecified gastrojejunal ulcer with hemorrhage with or without obstruction 534.6x Chronic or unspecified gastrojejunal ulcer with hemorrhage and perforation with or without obstruction 578.0 Hematemesis

	ICD-9 procedure code 44.43 Endoscopic control of gastric or duodenal bleeding CPT code 43255 Upper gastrointestinal endoscopy including esophagus, stomach and either the duodenum and/or jejunum as appropriate with control of bleeding, any method *Lower and unspecified G.I. bleeds:* 562.02 Diverticulosis of small intestine with hemorrhage 562.03 Diverticulitis of small intestine with hemorrhage 562.12 Diverticulosis of colon with hemorrhage 562.13 Diverticulitis of colon with hemorrhage 569.3x Hemorrhage of rectum and anus 569.85 Angiodysplasia of intestine with hemorrhage 578.1x Blood in stool 578.9 Hemorrhage of GI tract, unspecified *Other major bleeds:* 285.1x Acute posthemorrhagic anemia 423.0x Hemopericardium 459.0x Hemorrhage not specified 599.7 Hematuria 719.1x Hemathrosis 786.3x Hemoptysis 984.7x Epistaxis

Covariates	**Related ICD-9 or procedure codes**
Inclusion:	
Atrial fibrillation	427.31
Congestive heart failure	1 inpatient or 2 outpatient ICD-9 codes: 428.x 398.91 402.01 402.11 402.91 404.01 404.11 404.91 404.03 404.13 404.93
Past stroke	ICD-9 codes: 431.x 433.x 434.x 436.x 437.1 438.x
Past TIA	ICD-9 code: 435.x
Diabetes	1 hospital discharge or 2 outpatient ICD-9 for DM 250.x OR Dispensing of metformin, sulfonylureas, insulin, or other direct antidiabetic agent
Hypertension	ICD-9 diagnosis codes: 401.x-405.x

	OR Dispensing of CCB, ACEI, ARB, BB, thiazide diuretic, or other direct antihypertensive agent
Coronary Artery Disease (CAD)	At least 1 ICD-9 code from: 410.x-414.x 429.2 V45.81
Exclusion:	
Valvular heart disease and heart valve replacement	ICD-9 diagnosis codes: 394.x 395.x 396.x 397.x 398.9x V42.2 V43.3 ICD-9 procedure codes: 35.1x CPT codes: 33660-33665 33400-33403 33405 33420-33468 33420-33430 33460 33463-33468 33475

	33496 0257T 0258T 0259T 0262T
Active liver disease	ICD-9 diagnosis codes: 070.x 571.x-573.x 456.0-456.2x 155.0 155.1 155.2 576.8 ICD-9 procedure codes: 39.1 42.91
Cancer within the last 6 months	ICD-9 diagnosis codes: 140.x-208.x 230.x-239.x
Severe renal disease requiring dialysis	ICD-9 procedure codes: 39.95 54.98 56.0 V56.8 CPT codes: 90935-90993 99512 99559

Chronic renal insufficiency	ICD-9 codes: 582.x-583.x 585.x-587.x
Active or subacute endocarditis	ICD-9 codes: 421.1
Predicted probability of frailty > 15%	Calculated from Faurot et al.[31]
Other covariates	
Systemic embolism	ICD-9 codes: 444.x
Deep vein thrombosis	451.x, 453.x
Pulmonary embolism	415.11, 415.12, 415.19
Hyperlipidemia	ICD-9 codes: 272.0 272.2 272.4 OR Statins or other antihyperlipidemic
Atherosclerosis	ICD-9 codes: 440.9 414.x 429.2
Peripheral vascular disease	1 inpatient or 2 outpatient claims with the following codes: ICD-9 codes: 440.20-440.24 440.29-440.32 440.3 443.9 ICD-9 procedure codes: 38.08

	38.09 38.18 38.48 38.49 39.25 39.5 39.9 HCPCs: 35256, 35286, 35351, 35355, 35361, 35363, 35371, 35372, 35381, 35454, 35456, 35459, 35470, 35473, 35474, 35482, 35483, 35485, 35492, 35493, 35495, 35521, 35533, 35541, 35546, 35548, 35549, 35551, 35556, 35558, 35563, 35565, 35566, 35571, 35621, 35623, 35641, 35646, 35647, 35650, 35651, 35654, 35656, 35661, 35663, 35666, 35671
Acute renal disease (within the past month)	ICD-9 codes: 580.0 580.4 580.8 580.9 581.0 581.1 581.2 581.3 581.8 581.9 584.6 584.7 584.8 584.9
Smoking	ICD-9 code:

	305.1 649.0x 989.84
Obesity	ICD-9 code: 278.00
Alcoholism	ICD-9: 94.61-94.63 94.67-94.69 303.x 305.0x 291.x 357.5x 425.5x 571.1x 571.2x 571.3x

9.4 Data sources

Outcome Assessment in the RE-LY Trial

The RE-LY trial followed individuals for two primary outcomes after treatment initiation: first stroke or systemic embolism (efficacy) and first major hemorrhage (safety). These outcomes were reviewed and categorized by an international team of blinded adjudicators and patients were also provided symptom questionnaires at regular intervals; these symptom questionnaires were followed up on with medical record review. Stroke was defined as "sudden onset of a focal neurologic deficit in a location consistent with the territory of a major cerebral artery" and was divided into ischemic, hemorrhagic, and unspecified types. Major bleeding was defined as reduction in hemoglobin level of 20 grams per liter or more, transfusion with at least 2 units of blood, or symptomatic bleeding in a critical area or organ.

Within the context of our study, we will be using RE-LY's outcome data on first stroke, whether hemorrhagic, ischemic, or unspecified, and ignoring the systemic embolism outcomes (as it is difficult to compare these results given difficulties capturing embolisms in claims data)[32] with a secondary analysis focusing exclusively on ischemic stroke as an efficacy outcome; risk of hemorrhagic stroke may be elevated with a more potent anticoagulant while risk of an ischemic stroke is lowered. When assessing bleeding outcomes, we will be directly using the major bleeding outcomes from the trial, with a secondary analysis focusing specifically on gastrointestinal bleeding events. We will also use crude mortality data from the trial.

Exposure Assessment in the RE-LY Trial

In order to facilitate comparison with observational analyses and prevent the large gap in warfarin adherence from jeopardizing transportability, our main analyses will be an analysis censoring individuals at treatment switching or discontinuation. While there is potential for a per-protocol analysis censoring at treatment switching or discontinuation unless an individual experiences adverse events, it would be difficult to implement a comparable approach in observational data.[33]

In order to conduct any non-ITT analysis, we would have to be capable of tracking when individuals discontinued or switched the drug to which they were randomized. During the span of their follow-up visits, the RE-LY trial tracked individuals to see whether they discontinued the study drug. The authors did not appear to conduct any per-protocol or as-treated analyses in the study, so it is unclear exactly what data they will have available.

Given their regularly scheduled follow-up visits (14 days after randomization, at 1 month, at 3 months, at 6 months, at 9 months, at 12 months, and then every 4 months until the end of the study) it seems probable there is data on whether patients discontinued treatment at each of these time points, even if the exact date of discontinuation may not be not recorded. It is certain they have discontinuation data at the 1 and 2 year marks; the main publication reports that discontinuation varied slightly across study arms and was lower in the warfarin group than in the dabigatran groups at both one year (10% vs. 15%) and two years (17% vs. 21%) and only 20 patients (0.1%) were lost to follow-up across the entire study. If there is data on specific dates of discontinuation, we will explore giving an additional 1 week grace period to look for outcome events in the trial in secondary analyses. If there is only information about discontinuation at each follow-up visit, we will assume individuals discontinued at the midpoint of the interval in the main analysis and perform sensitivity analyses with discontinuation at the beginning or end of the interval as well.

In our intention-to-treat secondary analyses, exposure assessment for trial patients will be trivial: all their person-time will be assigned the exposure to which they were randomized.

Covariate Assessment in the RE-LY Trial

Two types of covariates will be assessed in RE-LY in addition to exposure and outcome: 1) effect measure modifiers for use in the sampling model and 2) variables associated with discontinuation and censoring. Fortunately, RE-LY collected a large quantity of baseline information on variables hypothesized to be associated with the outcome in order to examine treatment effects in various subgroups that can be leveraged in these analyses.

Effect measure modifiers: Potential effect measure modifiers were identified from causal diagrams depicting hypothesized causal relationships between sampling, our outcomes of

stroke and bleeding, and a variety of other variables. These diagrams were built by examining the marginal distributions of variables in trial and target populations (including in the literature review) as well as review of risk factors for stroke and bleeding for patients with AF. As mentioned in section of Background and Significance, a transport model that renders sampling independent of the outcome should allow estimation of an unbiased estimate in the target population. After adjusting for comorbidities, age, sex, frailty, and history of past medication use, the only path from sampling to either of our outcomes will be via the discontinuation and anagement node. Fortunately, censoring individuals at discontinuation or switching will reduce the potential for differing rates of discontinuation to be problematic, but without detailed lab and clinic data in both the trial and target population removing the potential for differences in warfarin management is unlikely to be possible. **Aim 1.1** is designed to assess the extent to which this or other unmeasured factors may bias our prediction of the outcome in the target population.

Censoring weight covariates: We will also use baseline and time-varying trial covariates to standardize the population continuing treatment to continue to be representative of the population at baseline even if some of the effect modifying covariates are associated with censoring (or, in this case, trial discontinuation). This is for two reasons: first, there is potential for introduction of confounding after initial randomization. If, for example, patients with hypertension are more likely to discontinue treatment with warfarin than treatment with dabigatran and hypertension is associated with increased stroke risk, an unweighted effect estimate censoring at discontinuation or switching would be biased in favor of dabigatran.[33] Inverse probability of discontinuation weights (or inverse probability of censoring weights) can help deal with this problem.

The second reason is particular to transporting effect estimates that requires continuous prescribing and deals with mismatched effect measure modifiers or risk factors in the two populations. Suppose that patients in the target population with diabetes in the target population are more likely to discontinue their warfarin or dabigatran, while patients in the trial population are not. If we do not weight or standardize both populations, we could see differences in our weighted trial effect estimate and observational population effect estimates

purely because at later time points the trial population includes more individuals with the effect measure modifier of diabetes. We could also see differences in our **Aim 1.1** assessment of transportability that arise only because of differences we've induced by requiring individuals to stay on their initial therapy.

The solution is stabilized inverse probability of censoring (sIPCW) weights. In the RE-LY trial, these will be estimated with stepwise logistic regression within each treatment arm assessing the probability of staying on treatment up to a given time point, then the next time point, then the next, and so on. Variables are assessed at baseline as well as at each time point (it is unclear whether RE-LY collected data on many effect modifiers after baseline). Individuals are then assigned weights according to the following equation where Z_i is the set of time-varying and baseline covariates associated with sampling and Tx is the treatment arm.

$$sIPCW = \frac{P(Censored_t = 0 |\ Censored_{t-1} = 0, Tx = X\ )}{P(Censored_{t=0}\ |\ Censored_{t-1} = 0,\ Z_i, Tx = X)}$$

We will explore separate analyses estimating these probabilities at quartiles of follow-up, quintiles of follow-up, and at 6 months, 12 months, and either 20 months (if specific dates of discontinuation are not available in the data) or 18 months (if discontinuation status is known at each follow-up visit). We will include effect measure modifiers as well as confounders that are known in the trial (discussed below in the observational section) in the model, as well as the last INR before discontinuation for patients in the warfarin arm should it be available. Fortunately, warfarin and dabigatran do not cause changes in many of these variables, particularly most of the comorbidities and age, so there are few mediators we have to worry about conditioning on.

Outcome Assessment in Medicare

Unfortunately, we do not have access to medical records for review by an international blinded group of adjudicators for use in this cohort. Instead, we will use ICD-9 codes in claims, both inpatient and outpatient, to identify stroke and major bleeding. We will use similar ICD-9 codes to identify the types of stroke that were used in past observational analyses using claims-based data to facilitate comparisons to their results (i.e., by Seeger et al.).[18] Medical record review

in some databases have shown positive predictive value of close to 90% for these codes, suggesting they perform quite well.[34] If we decide to use Medicare-linked North Carolina Data Warehouse data for computing inverse probability of censoring weights in patients from 2015 (see covariates below), we will also translate these ICD-9 codes to ICD-10 codes.
When identifying major bleeding events, we will also use ICD-9 codes in inpatient and outpatient claims. These definitions are similar to and adapted from the definitions used in the other observational studies in claims-based data, particular Seeger et al, and map directly to the trial outcomes.[18] These codes and definitions have been shown to have positive predicted values between 80% and 90% with medical chart review in claims databases, particularly in the setting of anticoagulant-associated adverse events.[35, 36] An additional analysis will be carried out separating gastrointestinal hemorrhage from other major bleeding. Again, these codes will be translated to ICD-10 codes if we extend the study window to take advantage of the Medicare to North Carolina Data Warehouse linkage. Mortality will be taken directly from the Medicare data.

Exposure Assessment in Medicare

Prescription claims data for Medicare beneficiaries will be used to identify warfarin and dabigatran initiators from 2010 to October 2015. Individuals will be defined as initiators if the days supply from their last prescription for an oral anticoagulant ran out at least 60 days prior to the initial prescription. We will perform sensitivity analyses with 90 and 180 day washout periods for identifying treatment initiation. In the analysis censoring at discontinuation individuals' follow-up time will be censored after switching medications as or having a 30-day gap in novel oral anticoagulant coverage or a 45 day gap in warfarin coverage, with the larger gap for warfarin provided due to the fact that pharmacy days' supply may be out of sync with the way patients are taking warfarin due to changes in directions for use at anticoagulation management appointments. There will be pill carry-over for dabigatran but not warfarin users for similar reasons. To help identify warfarin initiators who may start to pay purely out of pocket, CPT codes for INR draws (CPTs 85610, 99363, and 99364)[37] will "refresh" warfarin prescriptions and extend the length of follow-up for thirty days from the time of the CPT code.

Individuals will only be followed after their first initiation, even if they have multiple periods that qualify with a sufficient wash-out period. In the intention-to-treat analysis, individuals will be followed until death.

Covariate Assessment in Medicare

We will be assessing three kinds of covariates in the Medicare cohort: 1) effect measure modifiers, 2) confounders, and 3) censoring covariates.

Effect measure modifiers: Effect measure modifiers in the Medicare cohort will be assessed at the date of treatment initiation for the Medicare cohort and the date of switching to dabigatran or the date of matched warfarin treatment continuation in the prevalent user cohort. We will use a lookback period of one year in claims data to assess the presence of the variables in the prevalent user cohort, we'll identify covariates 1 year prior to the date of matching for the warfarin cohort, rather than 1 year prior to baseline**.** We may also explore the use of all-available lookback period approaches in sensitivity analyses as they may more directly parallel the assessment of history of hypertension in the RE-LY trial.

Confounding variables: Confounding variables will be assessed at the date of treatment initiation for the Medicare cohort and the date of switching to dabigatran or the date of matched warfarin treatment continuation in the prevalent user cohort, with the same one-year and all-available lookback approaches as the effect measure modifiers. The specific variables included are based upon a directed acyclic graph which indicates which variables form a minimally sufficient adjustment set in the Medicare cohort that close all open backdoor paths between treatment and the outcomes. Because the set of variables that affect treatment choice are identical with respect to both outcomes and stroke and major bleeding events share many risk factors, the same adjustment set will be used in each analysis. The set of variables can be defended into three main categories: demographics (age, sex, race, and socioeconomic status), comorbid conditions (transient ischemic attack, congestive heart failure, diabetes, smoking, weight, alcoholism, history of bleeding, and frailty), and past medication use (past warfarin use). The associations between the demographics, comorbidities, and outcome was built based upon literature review of various epidemiological studies and cardiovascular and

risk scores including the $CHADS_2$ score[38] and the Framingham Risk Score.[39] The associations between each of these variables and whether a patient might use dabigatran rather than warfarin came from questions regarding prescribing preferences to a medical professional working in an anticoagulation clinic, our literature review, and treatment guidelines.[5] As can be seen in the graph analyses requiring individuals to stay on treatment in this context involve removing a potential mediator of treatment effect (potential for discontinuing the treatment in question). We could use linked North Carolina Data Warehouse medical record data or linked Medicare Current Beneficiary Survey data to augment key confounders with poor capture in claims like weight and smoking using multiple imputation to estimate those distributions in the overall Medicare AF cohort in each of the analyses.

Censoring weight covariates: We will also build censoring weights using effect measure modifiers and confounders in the Medicare population for use in analyses condition on staying on initial treatment and ensure that the population that continues on treatment is standardized to look like the initial population in the Medicare target population; otherwise, as discussed in **Covariate Assessment in the RE-LY Trial,** selective dropout in this estimate could lead to either bias in internal validity (only healthy individuals stay on warfarin, while unhealthy users stay on both medication) or external validity (the target and trial populations change differently over time from dropout). Weights would be estimated separately within the Medicare population from the RE-LY population because the processes leading to discontinuation are likely quite different between the two populations; the discontinuation rates are much higher in claims than RE-LY.

9.4 Study size

We anticipate a cohort of approximately 157,000 Medicare beneficiaries based upon extrapolation from the FDAs matched study in the 100% Medicare sample.[15] They observed roughly 340,000 patients over a period of 26 months, while we will observe 20% of that quantity over a longer follow-up period of 60 months for a total of approximately 157,000 participants. We may observe fewer warfarin participants given historical warfarin trends over time, however, and we also have more stringent inclusion and exclusion criteria because of our attempts to mimic trial inclusion. Since the precision of the **Aim 1** and **Aim 3** analyses rely

primarily on the number of events in the RE-LY trial population, as well as their similarity to the Medicare target populations, the exact amount of power and precision of these analyses is unclear. Past transport analyses[40, 41] have shown somewhat less precision as the original trial with increases in the standard error after applying transport weights, but given the size of the original RE-LY trial treatment effects with respect to stroke, bleeding, and mortality this will still allow for reasonable inference and observation of whether effects are enhanced or attenuated. Most importantly, the **Aim 1** estimates will be the first attempt to leverage randomized clinical trial data on atrial fibrillation patients while still attempting to estimate a treatment effect in real-world Medicare patients with atrial fibrillation, making them the best way to accurately generalize to these larger populations.

Assuming a sample size of 157,000 Medicare beneficiaries will also provide sufficient power for the **Aim 2** analyses estimating treatment effects using only the observational data. Given the FDA study's stroke rate of roughly 12.6 events/1000 person-years, if we observe a similar average follow-up time of about 0.28 person-years per individual we would see a total of 43,960 person-years of follow-up and 553 stroke events, a slightly larger number than the FDA study. The other outcomes we will be investigating in **Aim 2** (bleeding and death) will be much more common than stroke, as shown by the bleeding analysis in the 5% Medicare sample.[16] These analyses will also be the ideal benchmark to compare the results of observational to weighted trial estimates because they will be calculated in the same target population as the **Aim 1** analysis.

9.5 Data management

Analyses will be conducted in SAS 9.4 for Window. Because the individual-level Medicare data must remain on their private server and the individual-level RE-LY trial data must remain on the CTDT server, we will have to adopt a unique data management plan for this project.
In order to construct the weighting models we will use to make the RE-LY trial and Medicare populations similar, we will have to consolidate the data in some fashion. Rather than try to combine the data on one platform and require combination of individual data, we propose exporting tables of the joint distributions of RE-LY trial patient characteristics that will be used

in the weighting models from the CSDR platform onto the outside server. These tables will simply be counts of the number of patients in the trial with each possible joint covariate distribution (i.e., one row will be the number of patients age 75 with hypertension, diabetes, congestive heart failure, kidney failure, a history of bleeding, past warfarin use, and any other effect modifiers; another row will be the number of patients with those characteristics but no past warfarin use; and so forth).These tables will not include birth dates, treatment, outcomes, treatment initiation dates, region, subsequent treatment adherence, or any identifiers. Only this aggregated data will be exported from the platform with no individual patient data being removed. We will combine this data with the Medicare data on the Sheps server to create the models we will use to estimate sampling probabilities for each covariate pattern. The coefficients of the betas in these models will then be returned to the CTDT server for use in constructing inverse odds of sampling weights for use in our two main analyses.

In order to perform our analyses comparing outcomes in the RE-LY trial to outcomes in the Medicare patients on dabigatran and warfarin (separately) we will need to bring data on the outcomes in Medicare patients onto the CTDT server. Because of concerns about individual-level outcomes that could be inferred from survival curve data, we will instead bring cumulative incidence step functions at each one-week interval. As soon as the project is complete, all Medicare-derived survival data will be removed from CTDT and all tabular RE-LY trial covariates will be removed from the Sheps server. In the interim, each data set will be kept only on secure servers and never on laptops or personal electronic devices.

9.6 Data analysis

We will 1) estimate 1 and 2-year risk differences comparing warfarin from NOACs in the Medicare AF population with transport methods and assess whether key assumptions have been met, 2) estimate 1 and 2-year risk differences using observational epidemiologic methods from only the Medicare AF population and 3) assess the impact of creating misclassified and missing data on effect measure modifiers using real data.

Aim 1.1: Assessing Transportability

Methods (initial analysis): The first step in this process is comparison of unadjusted risk of stroke, bleeding, and death in the individuals randomized to dabigatran in RE-LY to the unadjusted risk of each of the outcomes in the Medicare dabigatran patients (with a similar comparison for the warfarin patients in the RE-LY trial and Medicare). This comparison will be conducted both visually by overlaying the unadjusted Aalen-Johansen cumulative incidence curves and plotting the risk difference between the RE-LY trial and Medicare cohorts over time as well as quantitatively by computing risk differences at 6, 12, 18, and 24 months. Variances will be obtained with bootstrapping with 1000 replicates, bootstrapping populations before limiting to complete cases or performing subsample validation. For both groups, follow-up will be censored at discontinuation or switching in the main analysis and inverse probability of selection weights will be used to correct for potential bias; an intention-to-treat analysis will also be applied.

Methods (inverse odds of sampling weights): We will also assess whether weighting by measured effect measure modifiers remediates the disparity in outcomes between the groups, now relying on the export of count data on the joint distributions of various effect measure modifiers. We construct weights from a model assessing sampling into the RE-LY trial from the Medicare AF cohort within strata of exposure X (where X is either warfarin or dabigatran) in the target population; since exposure was randomized within the trial we can treat the trial cohort as having both exposures when building this model. Odds weights are necessary so that the trial population is not assumed to be part of the target population. If individuals are on treatment X in the RE-LY trial, they are weighted according to equation 4 with the probabilities being calculated from a logistic regression model, where X is either warfarin or dabigatran and EMM represents the set of all selected effect measure modifying covariates and interaction terms.

$$sIOSW = \left(\frac{P(RE-LY\ trial \mid X)}{1-P(RE-LY\ trial \mid X)}\right) * \left(\frac{1-P(RE-LY\ trial \mid EMM, X)}{P(RE-LY\ trial \mid EMM, X)}\right)$$

Individuals in Medicare will receive weights of 1. We will use quadratic restricted splines for age and three-way interactions between each term when fitting the logistic regression models to account for complex types of effect measure modification, provided there is enough data for it.

Other statistical considerations: We will repeat the survival analysis using these weights to construct the Aalen-Johansen cumulative incidence curves for the RE-LY trial and Medicare cohorts (separately for each anticoagulant) and compare them visually and quantitatively with risk differences at 6, 12, 18, and 24 months just as we did the crude cumulative incidence curves. Variances would be obtained with bootstraps of 1000 replicates, making sure to bootstrap both populations before taking complete cases or conducting any form of imputation or subsample validation. If the weighted trial and target population curves differ and there is a non-null treatment effect, it is impossible to have taken into account all modifiers that differ between the groups on both scales, making it less likely all modifiers of the risk difference have been controlled for and increasing the possibility that the assumptions necessary for transportability have been violated. These analyses will be particularly important to evaluate the performance of warfarin initiators in Medicare relative to warfarin patients in the RE-LY trial. We will also assess transportability after limiting ourselves to patients in the Medicare cohort with various predicted probabilities of frailty based on the frailty prediction algorithm created by Faurot et al.[31] to help eliminate potential for including individuals in our target populations that were not trial-eligible due to high frailty or short life expectancy. If these frailty-restricted Medicare populations have more similar outcomes to the RE-LY trial participants than the original Medicare cohort, **Aim 1.2** and **Aim 2** will use them as additional target populations.

Aim 1.2: Transporting Treatment Effects

Weighting methods: Normal inverse probability of treatment weights methods typically eliminate confounding by creating exposed and unexposed pseudo-populations with identical confounder distributions after estimating the probability of exposure condition on various covariates. With IOSW, we instead transform the RE-LY trial population into a pseudo-

population with an identical distribution of baseline effect measure modifiers as the Medicare AF population[24] by combining the joint categorical modifier RE-LY data and Medicare AF data and using logistic regression to identify the probability of selection based upon the selected effect measure modifiers. If individuals are enrolled in RE-LY, they receive stabilized inverse odds of sampling weights (sIOSW) based on this equation:

$$(sIOSW) = (\frac{P(RE-LY\ trial)}{1-P(RE-LY\ trial)}) * (\frac{1-P(RE-LY\ trial \mid EMM)}{P(RE-LY\ trial \mid EMM)})$$

Unlike in **Aim 1.1**, Medicare patients will receive weights of 0 instead of 1 and the estimates are no longer conditional on the specific type of treatment received; X has dropped out of the equation. If there are model convergence issues or weights become large when we have three-way interactions between most categorical terms (there is limited research on how large is too large in this transportability context; here, we will define large weights as single individuals having weights of > 5% of the population after stabilization) we will fit a model with two-way interaction terms and examine the marginal covariate balance using standardized mean differences between the target and RE-LY population.

Follow-up and censoring: To reduce potential issues with differential persistence between the trial and target population estimates influencing the comparison, we will primarily focus on an analysis censoring trial individuals at treatment switching or discontinuation. Stepwise inverse probability of censoring weights based upon the set of EMM and confounders as described in **Covariate Assessment in the RE-LY Trial** and fit separately in the trial and target populations for each treatment will be used to prevent potential selection bias in this approach. We will also explore an intention-to-treat design as a secondary analysis.

Other statistical considerations: The outcomes of interest (stroke, major bleeding, and death) will be compared on the risk difference scale by directly comparing the weighted survival curves in the RE-LY trial population (constructed with the above weights and Aalen-Johansen methods rather than assuming we can prevent competing risks[42]). The weighting process will be bootstrapped with 1000 iterations in order to generate 95% confidence limits for the risk differences, with both the trial and target populations being bootstrapped before any

imputation, limitation to complete cases, or subsample validation. This process will be repeated for each of the target populations of interest (warfarin initiators, dabigatran initiators, dabigatran initiators and switchers to dabigatran, and initiators of warfarin or dabigatran). If these estimates from **Aim 1.2** differ from the results of the RE-LY trial, it lends credence to the idea that there are meaningful effect measure modifiers in place that were differentially selected for by the trial enrollment process.

Aim 2.1: New User Active Comparator Design

Methods (inverse probability of treatment weights): We will balancing the confounding variables between dabigatran and warfarin initiators using inverse probability of treatment weighting. To use these weights, we will estimate the probability of dabigatran initiation based upon the confounders described above fitting two-way interaction terms and additional terms as required to achieve marginal standardized mean differences as small as possible, with any major confounders with an SMD greater than 0.1 being unacceptable (though this is a fairly arbitrary cut point).[43] Stabilized inverse probability of treatment weights (sIPTW) will be assigned based upon the model predicted probabilities.[44] Dabigatran initiators are weighted using equation 2 and warfarin initiators with equation 3, where Z_i represents the confounding variables used in the propensity score estimation.

$$\textit{Equation 2: } sIPTW_{Dabi} = \frac{P(dabigatran)}{P(dabigatran \mid Zi)}$$

$$\textit{Equation 3: } sIPTW_{Warf} = \frac{1-P(dabigatran)}{1-P(dabigatran \mid Zi)}$$

Methods (other statistical considerations): The outcomes of interest, 1 and 2-year stroke, major bleeding, and death will be compared by contrasting the IPTW and IPCW-weighted survival curves (constructed with Aalen-Johansen methods and the above weights for each individual) in the Medicare population at specific year intervals after applying sIPCW weights. We will bootstrap with 1000 replicates for estimating confidence intervals. Since sIPTW will estimate the treatment effect in the entire population and we would like treatment effect estimates in the dabigatran and warfarin users for comparison with the estimates obtained with sIOSW, we

will also use standardized mortality ratio (SMR)[45] weights to estimate treatment effects in dabigatran and warfarin users specifically.

Aim 2.2: Prevalent User Design

Methods (prevalent user design): In this design, time-conditional propensity scores taking into account time-varying covariates are used to match individuals switching from a traditional therapy to a newly marketed one with participants that have taken the traditional therapy for the same amount of time (or the same number of prescriptions) and continued on the traditional therapy with similar distributions of relevant confounders. If someone switches from warfarin to dabigatran after 180 days of warfarin follow-up, a person with the same number of days on warfarin who continued on warfarin is selected as their match. Because of the enormous heterogeneity in days supply across warfarin prescriptions, we will use the months elapsed since individuals first initiated warfarin rather than the absolute number of prescriptions. We will also experiment with "capping" the months since initiation when matching to warfarin users to switchers rather than searching for perfect matches, as most of the heterogeneity in risk of stroke and bleeding is likely in the initial time period when dosage is being frequently adjusted. We will also explore being increasing restrictive with respect to our definition of switching (requiring continuous new use of warfarin up until dabigatran initiation versus requiring new use of warfarin before dabigatran initiation vs requiring any use of warfarin before dabigatran initiation).

Methods (other statistical considerations): We will use similar bootstrapping, Aalen-Johansen, and sIPCW methods as in **Aim 2.1** to obtain 1 and 2-year risk differences in this target population as well.

Aim 2.3: Contrasting Treatment Effect Estimates

Methods: The estimates obtained under these **Aim 1** will be directly contrasted with the estimates obtained under the first approach in **Aim 2**. This process will be repeated for each of the four target populations: 1) new users of warfarin or dabigatran; 2) new initiators of

warfarin; 3) new initiators of dabigatran; and 4) the dabigatran initiators and switchers to dabigatran. Comparisons will be performed relatively (dividing **Aim 1** estimates by these **Aim 2** estimates), absolutely (subtracting the **Aim 1** estimates from the **Aim 2** estimate), and graphically using forest plots of point estimates and 95% confidence limits, with particular attention paid to shifts in estimate across the null between methods. If the results in **Aim 2.1** or **Aim 2.2** differ from those in **Aim 1**, either additional or incorrectly modeled confounding factors exist (e.g. socioeconomic status has an unblocked path to the outcome), additional or incorrectly modeled effect measure modifiers exist (e.g. better management in the trial than the target population), or both.

Aim 3 Approach

Overview: Missing data and misclassification can be problematic when it comes to the internal validity of randomized controlled trials and observational studies,[46-50] where they can cause anywhere from no bias at all to a large amount of bias. Little attention has been paid, however, to the problems that missing data and misclassification can potentially cause when generalizing or transporting treatment effects to target populations, though the difference in estimates between complete cases and multiply imputed cohorts has been assessed.[40] The current analysis presents an excellent opportunity to use real-world data in a case study where the type of missingness and misclassification can be manipulated by investigators. Missingness is more likely in the context of trials and prospective longitudinal cohort studies, while misclassification is more likely in the context of claims-based studies. An opportunity to study the ramifications of both at once would help further understanding of transporting both trial and observational studies to particular target populations.

Aim 3.1: Missing Data

Objectives: In **Aim 3.1**, we will identify the extent to which effect measure modifying covariates 1) missing completely at random (MCAR) 2) missing at random (MAR) or 3) missing not at random (MNAR) in a) the RE-LY trial or b) the Medicare population would alter the target population treatment effect estimates and their associated variance after applying the sIOSW

used in **Aim 1.2** and various techniques for dealing with missing data including complete case analysis, missingness weights, and multiple imputation.

Methods: We will first assume the RCT and Medicare complete case data are correctly classified, transport an absolute-scale treatment effect on stroke and bleeding risk, and use these results as the baseline. Alternatively, we could instead estimate relationships between all of the variables in the sampling and treatment effect models and re-simulate potential datasets. Regardless of the approach used, we would subsequently induce a variety of missingness scenarios and observe the resulting change in precision and estimate when using 1) complete case analysis, 2) missingness weights, and 3) multiple imputation. **Table 1** illustrates the specific scenarios that will be investigated with each of several variables (hypertension history, diabetes history, TIA history, and bleeding history), manipulating each variable to varying levels of missingness based upon the below equation describing a logistic relationship between the probability of missingness and various variables.

$$P_{missing}(Cov) = \frac{\log(\beta_{intercept} + \beta_{age} * age + \beta_{vitKantag} * vitKantag + \beta_{Cov} * Cov)}{(1 + \log(\beta_{intercept} + \beta_{age} * age + \beta_{vitKantag} * vitKantag + \beta_{Cov} * Cov)}$$

Table 1: Different scenarios examining the impact of missingness on bias and variance of target population risk differences.

Scenario #	Description of EMM* missingness	$e^{(MIssingness\ intercepts)}$	$e^{(MIssingness\ coefficients)}$
1	Baseline scenario	0	0
2	Missing completely at random in RE-LY	0.1; 0.3; 0.5	0
3	Missing completely at random in Medicare	0.1; 0.3; 0.5	0
4	Missing at random** in RE-LY	0.1; 0.3; 0.5	0.15, 1.5; 0.2, 2; 0.4, 4
5	Missing at random** in Medicare	0.1; 0.3; 0.5	0.15, 1.5; 0.2, 2; 0.4, 4

6	Missing not at random*** in RE-LY	0.1; 0.3; 0.5	0.15, 1.5; 0.2, 2; 0.4, 4
7	Missing not at random*** in Medicare	0.1; 0.3; 0.5	0.15, 1.5; 0.2, 2; 0.4, 4

*EMM = Effect measure modifier (hypertension history, diabetes history, transient ischemic attack history, bleeding event history)

**Missing inducers = age and past use of vitamin K antagonists, respectively

***Having the EMM will give you 1.5x, 2.0x, or 3.0x the odds of missing the variable

Here, $\beta_{intercept}$ refers to the log-odds for the baseline prevalence of missingness; β_{age} is the increase in log-odds of missingness for a one-unit increase in age; $\beta_{vitKantag}$ is the increase in log-odds of missingness if an individual has a history of use of vitamin K antagonists; and β_{Cov} is the increase in log-odds associated with the presence of the covariate in question.

In Scenario 1 in Table 1, for example, all the β coefficients will be equal to 0, as there will be no missingness at all. Scenario 2a would change the value of the $\beta_{intercept}$ term for RE-LY participants from 0 to 0.1; Scenario 2b would change it to 0.3; and Scenario 2c would change it to 0.5. Scenarios 3a through 3c would mirror this in Medicare population. In both of these cases, data on the covariate would be MCAR. Scenarios 4 and 5 begin to make the variables MAR by making variables missing dependent upon the other key modifiers of age and past use of vitamin K antagonists. Finally, Scenarios 6 and 7 introduce a non-zero value for β_{Cov} at which point the covariate is MNAR in RE-LY (Scenario 6) or Medicare (Scenario 7). The values were selected to represent a broad spectrum of potential associations and extent of missingness. Each of the scenarios will be repeated for each of the variables, as well as combinations in which missingness is incremental (e.g. only individuals missing variable 1 can be missing variable 2).

Missingness methods: Complete case analysis will be performed by removing all individuals with missing values for the variable after bootstrapping, then conducting the weighted analysis. Missingness weights will be computed using logistic regression and the variables listed as effect measure modifiers. Multiple imputation will also be conducted feeding the effect measure modifying variables in as the basis for the imputation.

Final steps: After missingness is dealt with, the same models and operations used in **Aim 1** and **Aim 2** will be used to estimate 1 and 2-year risk differences for stroke, bleeding, and death in the population of warfarin and dabigatran initiators, with 1000 bootstrapped replicates to estimate confidence intervals. These estimates will be compared with the risk difference estimates from the initial analysis and with one another graphically with forest plots and statistically by quantifying the difference in estimate and variance in each scenario.

Aim 3.2: Misclassified Data

Table 2: List of different scenarios examining the impact of misclassification on bias and variance of target population risk differences.

Scenario #	Description of EMM* misclassification	$e^{(\text{Misclassification intercepts})}$	$e^{(\text{misclassification coefficients})}$
1	Baseline scenario	0	0
2	Misclassified completely at random in RE-LY	0.1; 0.3; 0.5	0
3	Misclassified completely at random in Medicare	0.1; 0.3; 0.5	0
4	Misclassified at random in RE-LY	0.1; 0.3; 0.5	0.15, 1.5; 0.2, 2; 0.4, 4
5	Misclassified at random in Medicare	0.1; 0.3; 0.5	0.15, 1.5; 0.2, 2; 0.4, 4
6	Misclassified not at random*** in RE-LY	0.1; 0.3; 0.5	0.15, 1.5; 0.2, 2; 0.4, 4
7	Misclassified not at random*** in Medicare	0.1; 0.3; 0.5	0.15, 1.5; 0.2, 2; 0.4, 4

*EMM = Effect measure modifier (hypertension history, diabetes history, transient ischemic attack history, bleeding event history)

**Misclassification inducers = age and past use of vitamin K antagonists

***Having hypertension history (when misclassifying non-hypertension) or diabetes (when misclassifying hypertension) will give you 1.5x/2.0x/3.0x the odds of misclassification, and individuals with HTN (non-hypertension) or DM (hypertension) will have 0.5x/0.33x/0.25x the probability of being selected into the subsample.

Methods for misclassification: Just as in **Aim 3.1** the first step will be transporting an effect based upon either a complete case data set or some type of simulated results. Then we will induce a variety of misclassification scenarios in the same variables examined in **Aim 3.1** (described in **Table 2**), focusing on just one variable each time, and observe the resulting bias and change in precision when 1) there is no adjustment for misclassification or 2) a subset of the population has validation data that can be used to correct for some of the misclassification. Because we will be inducing misclassification in only one of the trial and target populations in each scenario, this misclassification will be differential with regards to sampling. The probability of misclassification will be calculated with an identical equation to the one used in **Aim 3.1.** These scenarios and the resulting equations parallel those in **Aim 3.1** until the final two scenarios in which data is rendered MiscNAR. Instead of the value of the variables increasing the possibility of misclassification, they'll instead rely on another variable (diabetes history for the analyses misclassifying hypertension; hypertension history for all other analyses) that is associated with selection for subsampling and is omitted from the subsample analysis and imputation.

Dealing with misclassification: In Scenarios 1-5, we will assign each individual a 1% probability of being selected for subsampling. In Scenarios 6 and 7, individuals without the "not at random" causing variable (hypertension for most analyses, diabetes for hypertension analyses) will have a 1% probability of being selected for subsampling while other individuals will have a differing probability based upon the subscenario. After subsampling, we will calibrate the variables in the full sample using imputation methods predicting the chance of misclassification given age and past vitamin K antagonist use. We may also experiment with additional scenarios where only individuals with a positive value for the covariate have a chance of misclassification (i.e.

perfect specificity with no false positives), as that may be more common in claims-based studies than misclassifying someone who doesn't have the covariate.

Final steps: After attempting to remedy the misclassification, we will perform the same analyses used in **Aim 1.2** to determine the change in estimate and its precision under each misclassification scenario and compare statistics the same way we did in **Aim 3.1**.

9.7 Quality control

We will not take any steps above and beyond those already conducted by the trial investigators (for the RE-LY trial data) and the maintainers of the Medicare data at the UNC Sheps Center (for the Medicare data). Both of these data sets are cleaned and secured independently. All statistical programming run on both types of data will be archived for potential re-use and availability upon request.

9.8 Limitations of the research methods

There are three central limitations to the proposed research. The first is that we cannot guarantee that the Medicare population is a nationally representative sample of the future population of older adults with AF in the United States, especially as comorbidity patterns change in our aging population. All inferences we make about transported effect estimates and whether or not the assumptions necessary for effect estimate transport have been met will refer specifically to the Medicare target population being considered, rather than some abstract notion of perfect transportability. Still, this population represents an important benchmark that is a major target for trials like RE-LY. If RE-LY's results cannot adequately transport to the Medicare population, it is unlikely results will be able to be transported to future groups of older adults. Moreover, this patient population is still an important population in which to understand the relative risks and benefits of dabigatran treatment.

The second central limitation involves the methods used in **Aim 1.1** to assess transportability. These methods are specific but not necessarily sensitive to situations in which key effect measure modifiers have not been measured. It is possible for the hazard ratio comparing warfarin patients in RE-LY to target population warfarin patients to be 1 even if there are still

important unmeasured effect modifiers on the scale we are interested in, as we can only compare one arm and modifiers may be "canceling" one another's effect on the risk. In addition, we cannot guarantee which scale or scales (multiplicative, additive, or both) the undetected factors are modifiers on, limiting our ability to say with certainty there are unmeasured modifiers on the scale we care about that differ between the two populations. If we have any interest in transporting both a ratio and absolute scale effect estimate, however, this limitation is ignorable. Even if we only care about one scale the analysis still encourages additional caution in describing accuracy of the effects we are estimating.

The third central limitation also involves the methods used in **Aim 1.1**: they only work in the setting of a non-null treatment effect in your target population. Given the results of past studies and the trial itself, this does not seem likely to be the case, but it will be important to interpret **Aim 1.1** results in light of this should **Aim 1.2** show a treatment effect centered on the null. In light of this fact, we plan to conduct **Aim 1.2** analyses for each target population regardless of the findings in **Aim 1.1**. If there are large gaps in outcome that raise concerns about transportability of non-null effects, they will be noted both during presentations and in any publications using information from **Aim 1.**

9.9 Other aspects

N/A

10. Protection of human subjects

This study was reviewed and approved by the University of North Carolina at Chapel Hill Institutional Review Board.

11. Management and reporting of adverse events

As we will not be collecting any new data on these participants, we will not be collecting any new adverse event data. However, if our study using the clinical trial data reveals any dramatically new findings, we will contact the appropriate individuals at Boehringer-Ingelheim.

12. Plans for disseminating and communicating study results

We anticipate three papers from this project for submission to major academic journals. The first will be based on Aim 2.1 and Aim 2.2 and estimate a treatment effect in Medicare target populations using propensity score methods and a new user active comparator and a prevalent new user design. The second, based on Aim 1, will document the methods generating the transported target population treatment effect while contrasting it with the observational estimate in the target population, alongside the analysis of differences in stroke and bleeding risk after adjusting for various risk factors. The third will describe the impact of rendering trial and target population data missing or misclassified on the transported treatment effect estimates. We also expect preliminary work from the first paper to be submitted to the 2019 International Society for Pharmacoepidemiology annual meeting for poster or oral presentation, with additional potential submission to the 2019 Society for Epidemiologic Research meeting. In addition to these papers and publications, the first analysis will become part of Michael Webster-Clark's dissertation work and be housed at UNC libraries.

13. References:

1. Rahman, F., G.F. Kwan, and E.J. Benjamin, *Global epidemiology of atrial fibrillation.* Nat Rev Cardiol, 2014. **11**(11): p. 639-54.
2. Savelieva, I., A. Bajpai, and A.J. Camm, *Stroke in atrial fibrillation: update on pathophysiology, new antithrombotic therapies, and evolution of procedures and devices.* Ann Med, 2007. **39**(5): p. 371-91.
3. Marini, C., et al., *Contribution of atrial fibrillation to incidence and outcome of ischemic stroke: results from a population-based study.* Stroke, 2005. **36**(6): p. 1115-9.
4. Rosand, J., et al., *The effect of warfarin and intensity of anticoagulation on outcome of intracerebral hemorrhage.* Arch Intern Med, 2004. **164**(8): p. 880-4.
5. Andrade, A.A., et al., *Clinical Benefit of American College of Chest Physicians versus European Society of Cardiology Guidelines for Stroke Prophylaxis in Atrial Fibrillation.* J Gen Intern Med, 2015. **30**(6): p. 777-82.
6. Connolly , S.J., et al., *Dabigatran versus Warfarin in Patients with Atrial Fibrillation.* New England Journal of Medicine, 2009. **361**(12): p. 1139-1151.
7. Desmaele, S., et al., *Clinical trials with direct oral anticoagulants for stroke prevention in atrial fibrillation: how representative are they for real life patients?* Eur J Clin Pharmacol, 2016. **72**(9): p. 1125-34.
8. Eichler, H.G., et al., *Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response.* Nat Rev Drug Discov, 2011. **10**(7): p. 495-506.
9. Tanislav, C., et al., *Baseline characteristics in stroke patients with atrial fibrillation: clinical trials versus clinical practice.* BMC Res Notes, 2015. **8**: p. 262.
10. Kennedy-Martin, T., et al., *A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results.* Trials, 2015. **16**.
11. Lip, G.Y., et al., *Major bleeding risk among non-valvular atrial fibrillation patients initiated on apixaban, dabigatran, rivaroxaban or warfarin: a "real-world" observational study in the United States.* Int J Clin Pract, 2016. **70**(9): p. 752-63.
12. Lip, G.Y., et al., *Real-world comparison of major bleeding risk among non-valvular atrial fibrillation patients initiated on apixaban, dabigatran, rivaroxaban, or warfarin. A propensity score matched analysis.* Thromb Haemost, 2016. **116**(5): p. 975-986.
13. Banerjee, A., et al., *Net clinical benefit of new oral anticoagulants (dabigatran, rivaroxaban, apixaban) versus no treatment in a 'real world' atrial fibrillation population: a modelling analysis based on a nationwide cohort study.* Thromb Haemost, 2012. **107**(3): p. 584-9.
14. Sturmer, T., et al., *Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution--a simulation study.* Am J Epidemiol, 2010. **172**(7): p. 843-54.
15. Graham, D.J., et al., *Cardiovascular, bleeding, and mortality risks in elderly Medicare patients treated with dabigatran or warfarin for non-valvular atrial fibrillation.* Circulation, 2014: p. CIRCULATIONAHA. 114.012061.
16. Hernandez, I., et al., *Risk of bleeding with dabigatran in atrial fibrillation.* JAMA Internal Medicine, 2015. **175**(1): p. 18-24.
17. Villines, T.C., et al., *A comparison of the safety and effectiveness of dabigatran and warfarin in non-valvular atrial fibrillation patients in a large healthcare system.* Thrombosis and haemostasis, 2015. **114**(6): p. 1290-1298.

18. Seeger, J.D., et al., *Safety and effectiveness of dabigatran and warfarin in routine care of patients with atrial fibrillation.* Thrombosis and Haemostasis, 2015. **114**(6): p. 1277-1289.
19. Larsen, T.B., et al., *Comparative effectiveness and safety of non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study.* BMJ, 2016. **353**: p. i3189.
20. Go, A.S., et al., *Outcomes of Dabigatran and Warfarin for Atrial Fibrillation in Contemporary Practice: A Retrospective Cohort Study.* Annals of Internal Medicine, 2017. **167**(12): p. 845-854.
21. Molteni, M. and C. Cimminiello, *Warfarin and atrial fibrillation: from ideal to real the warfarin affaire.* Thromb J, 2014. **12**(1): p. 5.
22. Najafzadeh, M. and S. Schneeweiss, *From Trial to Target Populations - Calibrating Real-World Data.* N Engl J Med, 2017. **376**(13): p. 1203-1205.
23. Hong JL, J.F.M., LoCasale R, Dempster S, Cole SR, and Stürmer T, *Generalizing Randomized Clinical Trials: Challenges Due to Missing Data on the Target Population. .* Pharmacoepidemiology and Drug Safety, 2016. **25**(Supplement): p. 23-24.
24. Cole, S.R. and E.A. Stuart, *Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial.* Am J Epidemiol, 2010. **172**(1): p. 107-15.
25. Stuart, E.A., C.P. Bradshaw, and P.J. Leaf, *Assessing the generalizability of randomized trial results to target populations.* Prev Sci, 2015. **16**(3): p. 475-85.
26. Westreich, D., et al., *Transportability of trial results using inverse odds of sampling weights.* Am J Epidemiol, 2017.
27. 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.* Journal of the American College of Cardiology, 2013. **61**(22): p. 2264-2273.
28. Curtis, L.H., et al., *Design considerations, architecture, and use of the Mini-Sentinel distributed data system.* Pharmacoepidemiology and Drug Safety, 2012. **21**(S1): p. 23-31.
29. Zhang, Z., *Survival analysis in the presence of competing risks.* Ann Transl Med, 2017. **5**(3).
30. Pearl, J. and E. Bareinboim, *External validity: From do-calculus to transportability across populations.* Statistical Science, 2014. **29**(4): p. 579-595.
31. Faurot, K.R., et al., *Using claims data to predict dependency in activities of daily living as a proxy for frailty.* Pharmacoepidemiol Drug Saf, 2015. **24**(1): p. 59-66.
32. Tamariz, L., T. Harkins, and V. Nair, *A systematic review of validated methods for identifying venous thromboembolism using administrative and claims data.* Pharmacoepidemiol Drug Saf, 2012. **21 Suppl 1**: p. 154-62.
33. Hernan, M.A. and J.M. Robins, *Per-Protocol Analyses of Pragmatic Trials.* N Engl J Med, 2017. **377**(14): p. 1391-1398.
34. Andrade, S.E., et al., *A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data.* Pharmacoepidemiol Drug Saf, 2012. **21 Suppl 1**: p. 100-28.
35. Wahl, P.M., et al., *Validation of claims-based diagnostic and procedure codes for cardiovascular and gastrointestinal serious adverse events in a commercially-insured population.* Pharmacoepidemiol Drug Saf, 2010. **19**(6): p. 596-603.
36. Cunningham, A., et al., *An Automated Database Case Definition for Serious Bleeding Related to Oral Anticoagulant Use.* Pharmacoepidemiol Drug Saf, 2011. **20**(6): p. 560-6.
37. Lauffenburger, J.C., et al., *Completeness of prescription information in US commercial claims databases.* Pharmacoepidemiol Drug Saf, 2013. **22**(8): p. 899-906.
38. Xing, Y., et al., *CHADS(2) score has a better predictive value than CHA(2)DS(2)-VASc score in elderly patients with atrial fibrillation.* Clin Interv Aging, 2016. **11**: p. 941-6.

39. Berry, J.D., et al., *Framingham risk score and prediction of coronary heart disease death in young men.* Am Heart J, 2007. **154**(1): p. 80-6.
40. Hong, J.-L., et al., *Generalizing Randomized Clinical Trial Results: Implementation and Challenges Related to Missing Data in the Target Population.* American Journal of Epidemiology, 2017: p. kwx287-kwx287.
41. Cole, S.R. and E.A. Stuart, *Generalizing evidence from randomized clinical trials to target populations the actg 320 trial.* American Journal of Epidemiology, 2010. **172**(1): p. 107-115.
42. Borgan, Ø., *Aalen–Johansen Estimator*, in *Encyclopedia of Biostatistics*. 2005, John Wiley & Sons, Ltd.
43. Austin, P.C., *Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples.* Stat Med, 2009. **28**(25): p. 3083-107.
44. Cole, S.R. and M.A. Hernan, *Adjusted survival curves with inverse probability weights.* Comput Methods Programs Biomed, 2004. **75**(1): p. 45-9.
45. Li, F., K.L. Morgan, and A.M. Zaslavsky, *Balancing covariates via propensity score weighting.* Journal of the American Statistical Association, 2017: p. 1-11.
46. Rothman KJ, G.S., Lash TL, editors., *Modern Epidemiology*. 3rd ed. 2008, New York, NY: Lippincott-Raven.
47. Daniel, R.M., et al., *Using causal diagrams to guide analysis in missing data problems.* Statistical Methods in Medical Research, 2012. **21**(3): p. 243-256.
48. Johnson, C.Y., et al., *Potential sensitivity of bias analysis results to incorrect assumptions of nondifferential or differential binary exposure misclassification.* Epidemiology, 2014. **25**(6): p. 902-9.
49. Nakasian, S.S., J.A. Rassen, and J.M. Franklin, *Effects of expanding the look-back period to all available data in the assessment of covariates.* Pharmacoepidemiol Drug Saf, 2017. **26**(8): p. 890-899.
50. Hernán, M.A. and S.R. Cole, *Invited Commentary: Causal Diagrams and Measurement Bias.* American Journal of Epidemiology, 2009. **170**(8): p. 959-962.

Annex 1. List of stand-alone documents

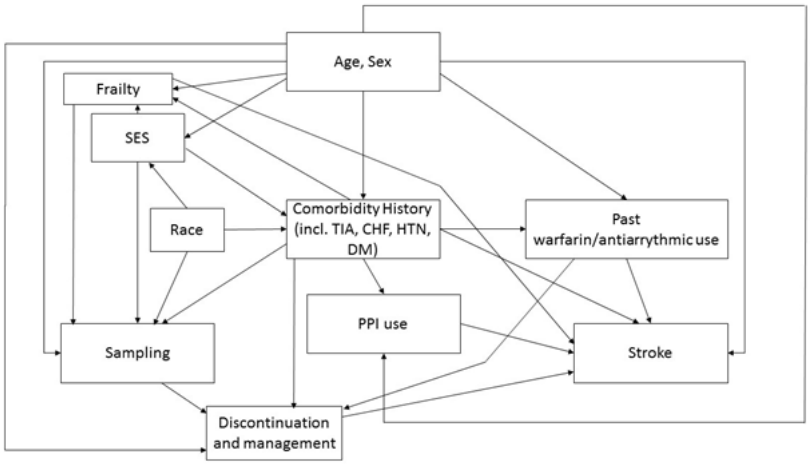
N/A

Annex 2. ENCePP checklist for study protocols

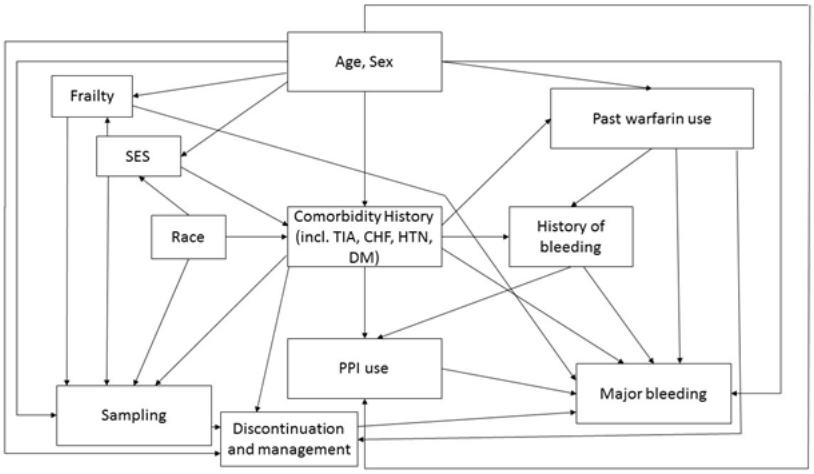
N/A

Annex 3. Additional information

Causal diagram for identifying potential effect modifiers of stroke:



Causal diagram for identifying potential effect modifiers of bleeding:



Causal diagram for identification of confounders: