1. Title Page

Title	TARGET-EU*:* Rivaroxaban and risk of major gastrointestinal bleeding in elderly patients with non-valvular atrial fibrillation
Research question & Objectives	To assess whether the risk of major gastrointestinal bleeding is different for new users of rivaroxaban vs. new users of apixaban aged 75 years or older with atrial fibrillation
Protocol version	V1.0
Last update date	27 February 2026
Contributors	**Primary investigator contact information:** Dr P.C. Souverein Dr D.L. Weir **Contributor names:** Patrick Souverein Daniala Weir
Study registration	**Site:** https://catalogues.ema.europa.eu/node/4440/administrative-details **Identifier:** EUPAS1000000539
Sponsor	**Organization:** EU PE&PV research network **Contact:** eupepv@uu.nl
Conflict of interest	None

Table of contents

2. Abstract

Background: Non-valvular Atrial fibrillation (NVAF) is a common arrhythmia that presents a significant morbidity and mortality risk. Current evidence suggests important differences in bleeding risk between DOACs that may be particularly relevant in high-risk population, such as the elderly, who tend to be underrepresented in RCTs. The purpose of this study is to compare the safety of apixaban versus rivaroxaban in older adults with NVAF, with a particular focus on major gastrointestinal bleeding.
Objectives: The primary objective is to estimate the effect of initiating rivaroxaban versus apixaban on time to a first major gastrointestinal bleeding event.
Methods: We will conduct an active comparator new-user cohort study using linked electronic health records from Denmark (Danish national registers) and Spain (SIDIAP). Eligible individuals are adults (≥75 years) who initiated rivaroxaban or apixaban between 2012 and 2023. In the primary analysis, a while on treatment strategy is used for treatment-related intercurrent events (discontinuation, switching). Inverse probability of treatment weighting (IPTW) is used to adjust for confounding. The primary analysis uses a Cox proportional hazards model, with supplemental analyses using an accelerated failure time model to estimate restricted mean survival time (RMST) at 1 and 2 years. Sensitivity analyses will be conducted to assess the impact of changing the conditions related to non-informative censoring using inverse probability of censoring weights as well as outcome misclassification using probabilistic bias analysis.

3. Amendments and updates

Version date	Version number	Section of protocol	Amendment or update	Reason
27 February 2026	1.0			

4. Milestones

Table 1. Milestones

Milestone	Date
Study protocol for RWD study	8 August 2025
Preliminary results RWD study	April 2026
Final Study report	10 June 2026

5. Background

What is known about the condition:

Non-valvular Atrial fibrillation (NVAF) is a common arrhythmia that presents a significant morbidity and mortality risk *[1]* Guidelines recommend Direct Oral Anticoagulants (DOACs) over Vitamin K Antagonists (VKAs) in patients with non-valvular AF due to several benefits of DOACs over VKAs, such as an improved efficacy/safety ratio, predictable effect without any need for monitoring, and fewer food and drug interactions. In geriatric patients with NVAF, DOACs tend to be underused or underdosed due to concerns of excessive fall-related intracranial bleeding, cognitive impairment, multiple drug-drug interactions, low body weight or impaired renal function.

What is known about the exposure of interest:

Although there are no clinical trials that specifically assessed the efficacy and safety of DOACs in older people, there were a considerable number of patients aged >75 years enrolled in the studies favouring their use. A sub-analysis of patients aged ≥75 years showed that all four DOACs have similar efficacy in preventing stroke and systemic embolism (SE) compared with warfarin [2]. A more recent patient-level meta-analysis of these four randomised controlled trials suggested that the relationship between age and major bleeding varied considerably between different DOACs. For example, there was an increased risk of GI bleeding with dabigatran and rivaroxaban among individuals aged >75 years; however, this was not observed with apixaban and edoxaban [3]. In a post-hoc analysis of the Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation (ROCKET AF), significantly higher gastrointestinal bleeding risks were found for rivaroxaban among patients aged ≥75 years or older, whereas similar major bleeding and intracranial bleeding risks for rivaroxaban as compared to warfarin were observed [6]. Similarly, in the Japanese J-ROCKET AF trial, rivaroxaban use in older AF patients was associated with a

similar major bleeding risk to warfarin [7]. In contrast, apixaban use has been associated with a significantly lower major bleeding and intracranial bleeding as compared to warfarin in AF patients ≥75 years old in a subgroup analysis of the ARISTOTLE trial [8]. For GI bleeding specifically, Garcia et al. found that that the event rate of major GI bleeding among patients in the ARISTOTLE trial without prior history of GI bleeding was 0.65/100 person-years [9].

Gaps in knowledge:

Despite these findings, older adults remain underrepresented in DOAC trials, and current evidence suggests important differences in bleeding risk between DOACs that may be particularly relevant in this high-risk population. The purpose of this study is to compare the safety of apixaban versus rivaroxaban in older adults with NVAF, with a particular focus on major gastrointestinal bleeding.

What is the expected contribution of this study?

We will perform an observational study using two Electronic Healthcare Record (EHR) databases to assess whether treatment there is difference in the risk of major gastrointestinal bleeding for rivaroxaban use in comparison to treatment with apixaban in patients with atrial fibrillation, thereby addressing the knowledge gap.

6. Research questions and objectives

The overall aim of this study is to determine whether the risk of major GI bleeding is higher with use of rivaroxaban compared to apixaban among people with non-valvular atrial fibrillation aged 75 years or older.

6.1. Primary Estimand 1

Research question answered by the estimand

What is the hazard ratio of major GI bleed for rivaroxaban use versus apixaban use for people with non-valvular atrial fibrillation aged 75 years or older while on treatment (i.e., before treatment discontinuation, switching) and while alive?

Table 2A. Estimand 1

Attribute	Target Trial	Target Trial Emulation	Comment
Population	Patients over the age of 75 with non-valvular atrial fibrillation	Patients over the age of 75 with non-valvular atrial fibrillation	Most coding system have codes for atrial fibrillation, but not specific for non-valvular atrial fibrillation. Valvular forms of atrial fibrillation will be identified using relevant codes for valve involvement and used for exclusion.
Treatment Conditions	Intervention- rivaroxaban (RIV) Control- apixaban (APX)	Initiation of RIV vs APX	Intention to initiate the study treatments (i.e., treatment allocation) will be emulated using the first observed prescription
Endpoint	Time to first major GI bleed	Same. Time to first occurrence of major GI bleed, defined using diagnostic codes in primary and secondary care	
Summary Measure	Hazard ratio	Hazard ratio	
Intercurrent events and strategies to handle them	Same for both treatment conditions Treatment discontinuation: while on treatment Treatment switch to another DOAC : while on treatment Switch to vitamin K antagonist: while on treatment Non-GI bleeding death: while alive	Intercurrent events are handled according to the same prespecified strategies of the hypothetical target trial. Treatment discontinuation is measured using prescription refill data where a gap of more than 90 days is considered as discontinuation. A treatment switch is defined as a prescription for a different DOAC than initiated or a VKA while not having discontinued.	Identification of the time of treatment discontinuation and switch are subject to inaccuracies due to the reliance on recorded prescribing data. The estimated end dates of prescriptions are based on assumptions about treatment duration or daily dose, and the estimation process involves uncertainty. This is an issue for the analysis since we are not interested in data after the occurrence of the IE.

Rationale for why selected strategies to handle intercurrent events are chosen

The risk of GI bleeding as a safety outcome is linked to current DOAC use, whereas past DOAC exposure would not be expected to increase the risk of bleeding events. It is assumed that the effect on GI bleeding of a DOAC treatment is sustained only while on that treatment (and not after stopping it). Therefore, we are interested only in follow-up before treatment discontinuation/switch, which implies the use of a "while on treatment" strategy for dealing with treatment discontinuation/switch intercurrent events.

Equally, we are interested in the risk of GI bleeding while patients are alive, thus the choice of the "while alive" strategy to handle the intercurrent event of death.

6.2. Supplementary Estimand 2

Research question answered by estimand 2

What is the difference in restricted mean survival time (at year 1 and year 2) to major GI bleed between rivaroxaban and apixaban in patients with non-valvular atrial fibrillation aged 75 years or older while on treatment (i.e., before treatment discontinuation, switching or new add-on antihyperglycemic therapy) and while alive?

Table 2B. Estimand 2

Attribute	Target Trial	Target Trial Emulation	Comment
Population	Patients over the age of 75 with atrial fibrillation	Patients over the age of 75 with atrial fibrillation	
Treatment Conditions	Intervention- rivaroxaban Control- apixaban	Initiation of RIV vs APX	Intention to initiate the study treatments (i.e.,treatment allocation) will be emulated using the first observed prescription
Endpoint	Time to first major GI bleed	Time to first occurrence of major GI bleed, defined using diagnostic codes in primary and secondary care	
Summary Measure	Difference in Restricted Mean Survival Time	Difference in Restricted Mean Survival Time	
Intercurrent events and strategies to handle them	Same for both treatment conditions Treatment discontinuation: while on treatment	Intercurrent events handled according to the same prespecified strategies	Identification of the time of treatment discontinuation and switch are subject to inaccuracies due to the reliance on

	Treatment switch to another DOAC : while on treatment Switch to vitamin K antagonist: while on treatment Non-GI bleeding death: while alive		recorded prescribing data. The estimated end dates of prescriptions are based on assumptions about treatment duration or daily dose, and the estimation process involves uncertainty. This is an issue for the analysis since we are not interested in data after the occurrence of the IE.

Rationale for why selected strategies to handle intercurrent events are chosen

The risk of GI bleeding as a safety outcome is associated with current use of DOACs and past DOAC exposure would plausibly not be associated with an increased risk of bleeding events. It is assumed that the effect on GI bleeding of a DOAC treatment is sustained only while on that treatment (and not after stopping it). Therefore, we are interested only in follow-up before treatment discontinuation/switch, which implies the use of a "while on treatment" strategy for dealing with treatment discontinuation/switch intercurrent events.

6.3. Supplementary Estimand 3

Research question answered by estimand 3

What is the hazard ratio of major GI bleed for rivaroxaban use versus apixaban use for people with non-valvular atrial fibrillation aged 75 years or older while alive and in the hypothetical scenario where treatment discontinuation or switching would not occur.

Table 2C. Estimand 3

	Estimand 3 **(Secondary Estimand)**	**Target Trial Emulation**	**Comment**
Population	Patients over the age of 75 with atrial fibrillation	Patients over the age of 75 with atrial fibrillation	
Treatment Conditions	Intervention- rivaroxaban Control- apixaban	Initiation of RIV vs APX	Intention to initiate the study treatments (i.e., treatment allocation) will be emulated using the first observed prescription.

Endpoint	Time to first major GI bleed	Time to first occurrence of major GI bleed, defined using diagnostic codes in primary and secondary care	
Summary Measure	Hazard ratio	Hazard ratio	
Intercurrent events and strategies to handle them	Same for both treatment conditions Treatment discontinuation: hypothetical strategy Treatment switch to another DOAC that is not in either treatment arm: hypothetical strategy Switch to vitamin K antagonist: hypothetical strategy Non-GI bleeding death: while alive	Same for both treatment conditions Treatment discontinuation: hypothetical strategy Treatment switch to another DOAC that is not in either treatment arm: hypothetical strategy Switch to vitamin K antagonist: hypothetical strategy Non-GI bleeding death: while alive	Identification of the time of treatment discontinuation and switch are subject to inaccuracies due to the reliance on recorded prescribing data. The estimated end dates of prescriptions are based on assumptions about treatment duration or daily dose, and the estimation process involves uncertainty.

Rationale for why selected strategies to handle intercurrent events are chosen

In estimand 3, we will apply a hypothetical strategy to handle treatment switching and discontinuation. This strategy is relevant for evaluating safety, in a hypothetical scenario in which all patients would remain exposed to treatment for the full trial duration—mirroring the intended use in clinical practice.

7. Research methods

7.1. Study design

Research design:

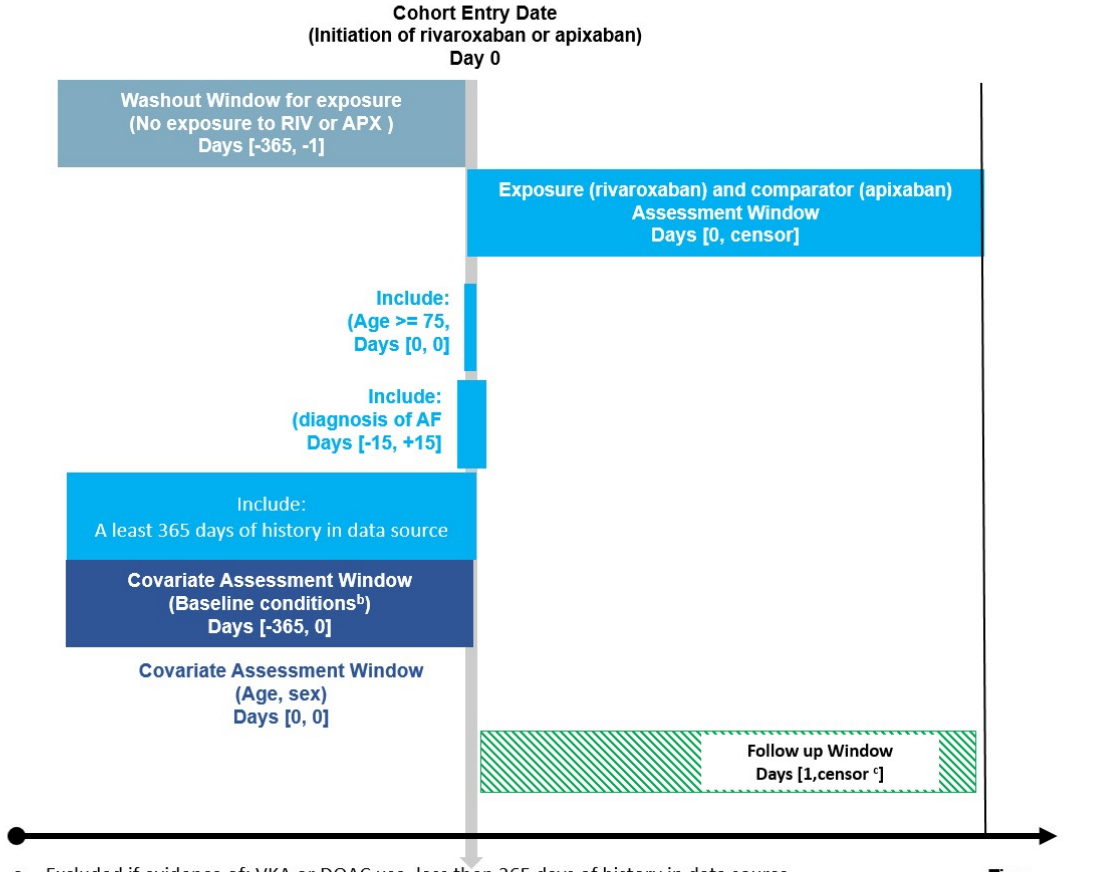
Active comparator new user retrospective cohort study

Rationale for study design choice:

Using an active comparator (rather than a non-user or general population) creates a more comparable reference group. Apixaban represents a clinically relevant and commonly prescribed alternative to rivaroxaban in daily clinical practice.

By comparing new users of these two drugs the design helps ensure that patients initiating either therapy have similar indications and disease severity at baseline, reducing confounding by disease severity or progression. The new-user design aligns the start of follow-up with treatment initiation, ensuring that all patients included are at risk of the outcome from a common time point. This avoids immortal time bias, which can occur if follow-up starts before the actual initiation of therapy.

7.2. *Study design diagram*



a. Excluded if evidence of: VKA or DOAC use, less then 365 days of history in data source
b. Full list and code algorithms provided in the published paper's appendix
c. Censored at earliest of outcome, death, disenrollment or end of the study period

7.3. Setting

This study is conducted using routinely collected electronic health records from 1 January 2013 to 31 December 2023, reflecting the period of routine clinical use of direct oral anticoagulants in older adults. The study is set within real-world healthcare systems in two European countries: Denmark and Spain. Data are sourced from the Danish national health registers and the Information System for Research in Primary Care (SIDIAP) in Catalonia.
The Danish registers provide nationwide, population-based coverage across all hospitals and outpatient settings, with linkage to prescription, hospital discharge, and mortality data. SIDIAP captures longitudinal primary care electronic health records covering approximately 78% of the Catalan population, with linkage to hospital discharge data and pharmacy invoicing records.

7.3.1 Definition of time 0 (and other primary time anchors) for entry to the study population

The definition of time 0 (index date of patients), is defined by a first prescription for rivaroxaban (intervention group) or apixaban (control group). This time 0 matches the date of randomisation in the hypothetical target trial. In the observational study, a stepwise approach will be applied as follows:

1) Identify a cohort of drug records for DOACs
2) The date of the first prescription of rivaroxaban or apixaban is the cohort entry date
3) Assess whether there is a diagnosis of NVAF around the time of the cohort entry date
4) Exclude subjects with history of VKA/others DOAC use in year before.

Table 3. Operational Definition of Time 0 (index date) and other primary time anchors

Study population name(s)	Time Anchor Description (e.g. time 0)	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type[2]	Diagnosis position	Incident with respect to	Measurement characteristics/ validation	Source of algorithm
Intervention cohort: rivaroxaban	Date of incident prescription for *RIV*	*Single*	*Incident*	*[-365,-1]*	*outpatient*	*ATC codes*	*n/a*	rivaroxaban or apixaban (any formulation of either)	*No validation study*	*n/a*
Control cohort: apixaban	Date of incident prescription for *APX*	*Single*	*Incident*	*[-365,-1]*	*outpatient*	*ATC codes*	*n/a*	rivaroxaban or apixaban (any formulation of either)	*No validation study*	*n/a*

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable

[2]See appendix for listing of clinical codes for each study parameter

7.3.2 Study inclusion criteria:

1. Initiate treatment with either rivaroxaban or apixaban: these are the therapies we want to compare
2. Age 75 years or older at index date: We are interested only in patients from that age onward.
3. A diagnosis of atrial fibrillation in a 30-day window around the index date: we are interested in patients receiving the intervention treatment or the control treatment for this indication only. As in daily clinical practice the recording of diagnosis records can be done retrospectively, some time after the date if treatment initiation is allowed.

Table 4. Operational Definitions of Inclusion Criteria

Criterion	Details	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Initiation of treatment	First of RIV or APX	[0]	OP	ATC		Intervention cohort: RIV Control cohort: APX		
Age ≥75yrs	Age >=75 years (year of index date-year of birth)	[0,0]	n/a	n/a	n/a	Intervention cohort: RIV Control cohort: APX	n/a	n/a
AF diagnosis	Diagnosis of AF	[-30, +30]	IP/OP	ICD-10	Any	Intervention cohort: RIV Control cohort: APX		

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.3.3 Study exclusion criteria

Exclusion criteria

Patients will be excluded if there is:

1) Less than 365 days of valid registration available prior to the index date
Rationale:
ensures sufficient time to assess new use of the drugs of interest and adequate time for assessment of covariates at baseline.

2) there is a diagnosis of valvular atrial fibrillation ever before the index date.
Rationale:
The indication for DOACs is non-valvular atrial fibrillation. Most coding systems only have atrial fibrillation as option, so we use this criterion to exclude those patients with e.g. prosthetic heart valves to exclude the valvular diagnoses.

3) There is use of DOAC or VKAs in the year prior to the cohort entry date.
Rationale: ensures that the cohort includes new users of either rivaroxaban or apixaban. By excluding individuals who have used either treatment class in the past year, the design emulates an incident user design, which aligns with the target trial framework and avoids biases associated with prior exposure (such as depletion of susceptibles or time-varying confounding by treatment history). It also allows for a clean assessment of the treatment effect.

Table 5. Operational Definitions of Exclusion Criteria

Criterion	Details	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Less than 365 days of history in data source	Assess whether there are at least 365 days of data availability prior to the index date	[-∞, 0]	All			All study populations		

Valvular Atrial fibrillation	Diagnosis of valvular atrial fibrillation any time before the index date	[-∞, 0]	OP, IP	ICD-10DA/Snomed	Any	All study populations		
History of DOAC/VKA use	Any prescription for any DOAC or VKA in the year before the index date	[-365, -1]	OP, OT	ATC	Any	All study populations		

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4. Variables

7.4.1 Exposure(s) of interest

The exposure group consists of patients who initiate rivaroxaban at any dose, formulation, or regimen. Treatment with rivaroxaban may last for up to two years. The comparator group consists of patients who initiate apixaban at any dose, formulation, or regimen. Treatment with apixaban may also last for up to two years. This amount of time should be sufficient, as the incidence of major GIB is highest in the first 6 months post-initiation. The use of an active comparator is the most appropriate approach because it aligns with real-world clinical practice and helps to address confounding by indication. It allows for the safety of rivaroxaban to be benchmarked against another medication that is commonly prescribed to treat patients with NVAF.

Algorithm to define duration of exposure effect:

For the construction of treatment episodes based on prescription records, the duration of each prescription is first derived using information available in the database. When end date or days supply is recorded, this information is used to directly define the prescription duration along with the date of the prescription. If days supply is not recorded, prescription duration will be estimated based on the quantity prescribed and dosing instructions. If dosing information is missing or incomplete, standard dosing assumptions consistent with routine clinical practice will be used (e.g., one defined daily

dose according to WHO ATC classification). Overlapping days between prescriptions are handled by carrying forward any unused supply. Specifically, if a refill occurs before the end of the previous prescription's calculated days' supply, the overlapping days are added to the end of the new prescription's duration. We assume that each dispensed prescription has a lasting effect of up to 90 days. Therefore, a gap of up to 90 days between the end of a prescription's days' supply and the subsequent refill is allowed, without considering the patient as having discontinued treatment. The choice of a 90-day window aligns with common prescribing and dispensing practices for chronic medications such as glucose-lowering therapies and is consistent with prior pharmacoepidemiologic studies. This grace period accommodates typical variations in refill timing, medication stockpiling, or short treatment interruptions, thereby reducing the risk of misclassifying ongoing therapy as discontinued. Additionally, for the final prescription in a treatment episode, we extend exposure by 90 days beyond the calculated end date to account for any residual pharmacological effect or continued use.

Table 6. Operational Definitions of Exposure

Exposure group name(s)	Details	Washout window	Assessment Window	Care Setting[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Incident with respect to…	Measurement characteristics/ validation	Source of algorithm
Exposure	Rivaroxaban	[-365, -1]	[0, censor]	OP	ATC	n/a	Exposure: Rivaroxaban	DOAC Prescriptions measured in one year prior to first use	No validation study	Investigator review of generic names
Comparator	Apixaban	[-365, -1]	[0, censor]	OP	ATC	n/a	Comparator: Apixaban	DOAC Prescriptions measured in one year prior to first use	No validation study	Investigator review of generic names

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4.2 Outcome(s) of interest

The primary endpoint will be time to the first occurrence of a major gastrointestinal bleeding. The definition of what is major bleeding according to the International Society on Thrombosis and Haemostasis (ISTH) is a fatal bleeding or a symptomatic bleeding in a critical area or organ, including bleeding in locations such as the retroperitoneum, central nervous system (intracranial or intraspinal), intraocular region, or joints (intra-articular), significant drop

in haemoglobin level decrease of at least 2 g/dL (or 20 g/L) or the need for a blood transfusion of at least two units of packed red blood cells. In RWD sources, the latter two criteria are difficult to capture. [11] Hence, the first occurrence of a relevant diagnosis code will be used to identify outcome events.

Table 7. Operational Definitions of Outcome

Outcome name	Details	Primary outcome?	Type of outcome	Washout window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source of algorithm
Major GI bleeding	Diagnosis of a major GI bleeding identified through hospital admissions or primary care records or death records	yes	Time-to-event	[-365-0]	IP, OP	-ICD10, ICD10 Danish version, ICD10CM	Primary	Exposure: rivaroxaban; comparator Apixaban	Same code list for GIB as used in earlier projects for consistency reasons	

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4.3 Follow up

Patients will be followed up from the date of treatment initiation until the outcome of interest, the end of the study period, the end of data collection or death of the patient, whichever comes first.

Table 8. Operational Definitions of Follow Up

	Select all that apply	Specify
Follow up start	Day 1	
Follow up end[1]		
Date of outcome	Yes	First occurrence of major GI bleeding during follow-up
Date of death	Yes	Administrative censoring at date of non-GI bleeding death
End of observation in data	Yes	Censor at date of last healthcare contact, known deregistration, (non-administrative censoring)
Day X following index date *(specify day)*	Yes	Day 730 (2 years after index date); administrative censoring
End of study period (specify date)	Yes	31-12-2023; date of administrative censoring
End of exposure *(specify operational details, e.g. stockpiling algorithm, grace period)*	Yes	Used to define follow-up end for estimand 1 and 2 (while on treatment effects) and end of observed follow-up of interest estimand 3 (f hypothetical approach)
Date of add to/switch from exposure *(specify algorithm)*	Yes	Used to define follow-up end for estimand 1 and 2 (while on treatment effects) and estimand 3 (f hypothetical approach)

[1] Follow up ends at the first occurrence of any of the selected criteria that end follow up.

7.4.4 Covariates (confounding variables and effect modifiers, e.g. risk factors, comorbidities, comedications)

We will identify demographics, comorbidities, comedications and other clinical variables that are risk factors for GI bleeding. Important risk factors considered for GI bleeding were a prior history of GI bleeding, peptic ulcer disease, diverticular disease, inflammatory bowel disease (IBD), chronic liver disease, gastroesophageal reflux disease (GERD), gastritis/oesophagitis, oesophageal varices, chronic kidney disease (CKD), hypertension, congestive heart failure, presence of GI malignancies, COPD, alcohol abuse, as well as concomitant use of medicines that modify haemostasis or increase the gastrointestinal bleeding risk such as nonsteroidal anti-inflammatory drugs, corticosteroids, selective serotonin reuptake inhibitors, antiplatelet drugs, PPIs. Furthermore, we will assess lifestyle factors (smoking, BMI).

Table 9. Operational Definitions of Covariates

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Demographics							Exposure: Rivaroxaban, Comparator: Apixaban		
Age at index date		Continuous	[0,0]			Any			
Gender		Binary	[0,0]			Any			
Comorbidities	At least one record in history			*IP, OP, OT*	*ICD-10, ICD-10 CM, ICD10 Danish version*	*Any*	Exposure: Rivaroxaban, Comparator: Apixaban		
Cancer GI tract		Binary	[-365,0]						
History of GIB		Binary	[-infin,0]						
History of GIB in year before cohort entry		Binary	[-365,0]						
Number of days (or other time unit) between most recent GIB diagnosis and index date		Continuous	[-infin,0]						
Peptic ulcer disease		Binary	[-365,0]						
Inflammatory bowel disease		Binary	[-infin,0]						
Gastroesophageal reflux disease (GERD)		Binary	[-infin,0]						
gastritis/esophagitis		Binary	[-365,0]						

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Oesophageal varices		Binary	[-infin,0]						
Chronic Kidney Disease stage 1-3		Binary	[-infin,0]						
Chronic Kidney Disease Stage 4-5		Binary	[-infin,0]						
Hypertension		Binary	[-infin,0]						
Chronic liver disease		Binary	[-infin,0]						
COPD		Binary	[-infin,0]						
Alcohol abuse	diagnostic codes for alcohol abuse related comorbidities will be searched	Binary	[-infin,0]						
Comedication	At least one prescription in time window				***ATC***		Exposure: Rivaroxaban, Comparator: Apixaban		
Corticosteroids		Binary	*[-182,0]*						
NSAIDs		Binary	*[-182,0]*						
SSRIs		Binary	*[-182,0]*						
PPIs		Binary	*[-182,0]*						
Antiplatelet drugs		Binary	*[-182,0]*						

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Lifestyle factors							Exposure: Rivaroxaban, Comparator: Apixaban		
Smoking	Most recent status	Binary	[-365,0]	OP, IP					
Body Mass Index	Most recent		[-365,0]	OP, IP					

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.5 Core Emulation Table – Design summary

Table 10. Comparison of Target Trial and Proposed Target Trial Emulation Design Elements

	Target Trial	Target Trial Emulation	Comment
Inclusion criteria	Age ≥ 75 Diagnosed with NVAF	Age ≥ 75 NVAF in 30 days window around the initiation date of treatment Treatment initiation with either RIV or APX	Eligibility applied using structured EHR data. Absence/misclassification of NVAF diagnosis can lead to smaller sample size. NVAF is established based on diagnosis of AF and the absence of evidence of valvular conditions Emulation restricts to new users in routine care

Exclusion criteria	Prior use of DOACs or VKAs within the last year prior to randomisation History of valvular atrial fibrillation	Same	Operationalised using prescription and diagnostic codes; will require lookback windows for classification
Setting	Multicenter	Routine care data sources (e.g., primary care and/or administrative databases) capturing prescriptions, outcomes, and covariates	Real-world data captures care as delivered
Time *(when follow up begins and ends):*	Begins at randomisation; Ends at first occurrence of outcome, treatment discontinuation, treatment switch, study withdrawal, loss to follow-up, non-GI bleeding death or at 2 years after randomisation	Begins at treatment initiation which is first prescription of RIV or APX; Ends at first outcome occurrence, treatment discontinuation, switch, non-GI bleeding death or at 2 years after treatment initiation, exit of individual within data source	Aligns start of follow-up with treatment initiation to mimic start of trial;
Study treatment conditions	Rivaroxaban vs. apixaban, both added to usual care; real-world use without restriction	Initiation of Rivaroxaban or apixaban; real-world use without restriction	
Outcome (including operational definition)	Time to first major GI bleeding, including death due to GI bleeding	Same outcome identified using diagnostic records in linked databases	Code lists and outcome definitions validated or informed by prior emulations. Cause of death might not be well recorded in SIDIAP
Method of Assignment to Trial Intervention	Simple 1:1 randomization	Assignment reflects clinical need. Inverse probability of treatment weighting (IPTW) will be used to adjust for baseline confounders.	Randomisation cannot be directly emulated. IPTW will be used in the statistical analysis to balance measured confounders in absence of randomisation.

Intercurrent Events and strategies to handle them.	Treatment discontinuation: while on treatment Treatment switch: while on treatment Non-GI bleeding death: while alive	Same strategies implemented based on prescribing & mortality data and using administrative censoring (or lack of for these intercurrent events) **Operational definitions:** • Treatment discontinuation is identified using prescription refill data, where a gap of more than 90 days between refills is considered a discontinuation. • Treatment switching is similarly measured using prescription records, with a gap of more than 90 days and receipt of another DOAC or VKA indicating a switch to a new therapy.	Identification of treatment discontinuation and switch time will be a limitation Cause of death may not be accurately identified in all cases, therefore, non-GI bleeding death may be misclassified
Loss to follow up	Patients who fail to return for the required study visits and his/her health condition and vital status remains unknown despite multiple attempts to contact them.	Patients with known de-registration date or database end. This is directly measured in RWD source.	Real-world proxy used to define loss to follow-up;

Estimand 2 (while on treatment) and 3 (hypothetical strategy; follow-up ends at the occurrence of each intercurrent event) are similar.

7.6 Data analysis

7.6.1 Analysis plan

Overview

The analyses will be conducted within a target trial emulation framework to estimate the comparative effect of the treatments of interest on time to the first occurrence of the outcome.

For Estimand 1, the primary causal effect measure is the hazard ratio, estimated using an inverse probability of treatment–weighted (IPTW) Cox proportional hazards model. The primary analysis will be performed separately within each data source, and the resulting hazard ratio estimates will be pooled using a random-effects meta-analysis. Potential sources of heterogeneity between data sources will be described qualitatively, including structural differences (e.g. coding systems, population coverage), measurement differences (e.g. recording practices), and their potential implications for the analysis (e.g. residual confounding or misclassification).

A comprehensive set of supplementary analyses will be conducted to support and contextualize the primary findings. These include crude and IPTW-adjusted Kaplan–Meier curves, crude Cox models, event counts and incidence rates, diagnostics of propensity score and weight distributions, covariate balance before and after weighting, assessment of censoring and intercurrent event patterns, proportional hazards assumptions, and evaluation of positivity.

Sensitivity analyses will be performed to assess the robustness of the primary results to key assumptions, including alternative approaches to handling censoring (e.g. IPCW) and bias analyses addressing potential exposure misclassification or other sources of uncertainty.

7.6.2 Primary Estimand (1) Analysis

i. Objective

To estimate the effect of initiating rivaroxaban versus apixaban on time to a first major gastrointestinal bleeding event among older adults with non-valvular atrial fibrillation. The analysis will focus on estimating the causal contrast without formally testing a pre-specified hypothesis.

ii. Exposure contrast

Rivaroxaban vs. Apixaban

iii. Outcome

Time to first major gastrointestinal bleeding event

iv. Analytic software:

R

v. Handling of intercurrent events

Handle intercurrent events according to the following strategies:
- Treatment discontinuation: Apply a while on treatment strategy
- Treatment switching: Apply a while on treatment strategy
- Non-GI bleeding death: while alive

Since while alive and while on treatment approaches are used, follow-up data after the occurrence of intercurrent events is not of interest and will not be analysed. Censoring at these IE is considered “administrative”

vi. Outcome Modelling

A Cox proportional hazards model, weighted by inverse probability of treatment (IPTW), will be used to estimate the effect of initiating rivaroxaban versus apixaban on time to first major GI bleeding. Crude Kaplan–Meier cumulative incidence curves will be estimated separately for patients initiating rivaroxaban and for those initiating apixaban . We will also estimate inverse probability of treatment weighted (IPTW) adjusted Kaplan Meier curves; each patient will be weighted by the inverse probability of their propensity scores.

- Start of follow-up: date of treatment initiation for either RIV or APX
- Endpoint: Time from treatment initiation to the first occurrence of major GIB.
- Censoring:
  - Non-administrative: Censoring due to loss to follow-up.
  - Administrative: Treatment discontinuation, Treatment switch, Non-GI bleeding death, database end, Censoring at end of study follow-up in the absence of GIB (maximum 2-years).
- Model covariate: Treatment group (RIV vs. APX)

vii. Confounding Adjustment

Inverse Probability of Treatment Weighting (IPTW)

To adjust for baseline confounding, inverse probability of treatment weighting will be used. Propensity scores, defined as the probability of initiating rivaroxaban versus apixaban, will be estimated using logistic regression. The model will include baseline covariates selected a priori based on clinical relevance and prior evidence as potential confounders—specifically, variables considered to be associated with the outcome, plausibly associated with treatment but not affected by treatment initiation. Variables associated with the treatment alone or variables which are a common effect of the exposure and another variable that influences the outcome will be excluded. These include patient demographics, lifestyle factors, comedications and comorbidities. A directed acyclic graph (DAG) will inform the conceptual identification of confounding structures, although it is not formally presented here. Restricted cubic splines will be used to model continuous variables (e.g., age).

Stabilized weights will be calculated by dividing the marginal probability of receiving the treatment actually received (i.e., the overall proportion treated in the study population) by the individual's estimated propensity score (i.e., the conditional probability of receiving their observed treatment). Weights will be truncated at the 1st and 99th percentiles to limit the influence of extreme values.

Weight truncation reduces the influence of individuals with highly improbable treatment assignments but does not resolve propensity score non-overlap. Therefore, if regions of the propensity score distribution show insufficient overlap, we plan to restrict analyses to the overlapping region (trimming) or apply overlap weights.

Truncated stabilized IPTW weights will then be applied in the Cox proportional hazards model (weighted likelihood) to estimate the marginal treatment effect (rivaroxaban vs apixaban on time to first GI bleed). Standard errors will be estimated using robust (sandwich) variance estimators.

Assumptions Underlying IPTW

- **No unmeasured confounding** (all relevant baseline confounders are included in the propensity score model).
- **Positivity** (each individual has a non-zero probability of receiving either treatment, given their covariates).
- **Correct model specification (**the propensity score model is correctly specified [functional form, covariate inclusion]).
- **Consistency (**each individual's potential outcome under the observed treatment equals their actual outcome).

Diagnostics for IPTW

- **Covariate balance**: Check that baseline characteristics are balanced across treatment groups after weighting.
    - Evaluate standardized mean differences (SMDs): SMDs < 0.1 will be considered acceptable.

Positivity check: Ensure adequate overlap in propensity score distributions between treatment groups to support estimation (graphically). Figures representing the distribution of IPTW for each treatment arm (to be conducted before and after truncation).

viii. Missing Data handling

Missing Exposure Data
We assume that missing refill or prescription records for rivaroxaban or apixaban reflect true treatment discontinuation after 90 days, and not incomplete data capture or prescriptions issued outside the database

Missing Outcome Data
The Cox proportional hazards model implicitly assumes non-informative censoring, meaning that censored participants contribute time at risk up to the time of censoring and their censoring is unrelated to the outcome, **conditional** on model covariates (the same covariates used to estimate the inverse probability of treatment weights) and not having experienced major GI bleeding up to the time of censoring. (i.e., outcome data is missing at random under these assumptions). In addition, IPTW may partially mitigate bias due to informative censoring insofar as censoring depends on these measured covariates.

Missing Covariate Data
- The absence of a disease diagnosis code is assumed to indicate the absence of the corresponding condition. This assumption may introduce potential misclassification if the condition exists but is not recorded in the data. A similar assumption applies to outcome variables, whereby the absence of an endpoint event during the available follow-up is interpreted as non-occurrence of the outcome.
- Missing values for lifestyle (smoking status) and BMI will be addressed using multiple imputation with chained equations (MICE) under the missing at random assumption.

Assessment of Missingness
Before performing imputation, we will examine the extent and patterns of missingness to evaluate whether imputation is appropriate. Specifically, we will:
- Quantify the percentage of missing data for each covariate.
- Compare the proportion of missing values across treatment groups to assess differential missingness.
- If a covariate has more than 40% missing data, we will consider alternative approaches (e.g., exclusion of the variable, sensitivity analyses) and justify the decision. Thresholds of 40% have been cited because effect estimates begin to be less reliable as the level of missingness increases beyond this threshold. [Jakobsen et al. (2017), “When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts”]

Imputation Model
MICE models will include all covariates used in the outcome and treatment models, as well as predictors of missingness and the exposure and outcome of interest. The treatment and outcomes of interest will also be included. Key covariates included in the imputation model will be:
- Demographics (age, sex)
- Clinical history and comorbidities
- BMI
- Lifestyle factors (e.g., smoking)

- Medication use

Full Conditional Distributions
MICE will use variable-specific conditional models:
- Logistic regression for binary variables (e.g., smoking yes/no).
- Multinomial logistic regression for categorical variables with >2 categories.
- Predictive mean matching for continuous variables

Number of Imputations and Diagnostics
We will generate at least 10 imputed datasets (to ensure stable estimates given the level of missingness) and pool results across imputations using Rubin's rules. Diagnostics will include:
- Checking whether imputed values are plausible and consistent with observed distributions.
- Evaluating convergence of the chained equations.
- Assessing stability and consistency of results across imputed datasets.

Effect Estimation Under Multiple Imputation
The imputation model will be applied prior to IPTW and effect estimation. IPTW and outcome models will then be fitted in each imputed dataset, producing treatment effect estimates (i.e., hazard ratios) and corresponding variances. These estimates will be combined across the imputed datasets using Rubin's rules, which account for both within-imputation variance (the average estimation error within each imputed dataset) and between-imputation variance (the variability in estimates across imputations). The total variance therefore reflects uncertainty from both the imputation process and the effect estimation, producing valid confidence intervals.

7.6.3 Supplemental Estimand (2) Analysis

Same as 7.6.2, except outcome model is IPTW weighted Accelerated Failure Time (AFT) model, followed by estimation of Restricted Mean Survival Time (RMST) at fixed time points (1 and 2 years).

Model Assumptions:
- Survival times follow a Weibull distribution
- Non-informative censoring (conditional on included covariates and survival up to time)
- Log-linear relationship between covariates and log survival time

To estimate the RMST at1 and 2 years from the Weibull AFT model, we first use the model to obtain the predicted survival curve for each treatment group. The RMST is then calculated as the average survival time up to a fixed time point, which corresponds to the area under the survival curve between time zero and the chosen time horizon 1 or 2 years).

- Fit the Weibull AFT model, which gives the shape and scale of the survival curve for each group.
- From this model, generate the predicted survival probability at each time.
- Integrate (i.e., add up) the survival probabilities from time 0 to 2 years and separately from time 0 to 2 years. The result is the expected survival time lived within those windows.
- Compare the RMST values between treatment groups to obtain the difference in average survival time over 1 and 2 years.

7.6.4 Supplemental Estimand (3) Analysis

Same as primary estimand but intercurrent events are handled using a hypothetical strategy. Subjects censored at the time of intercurrent event are non-administratively censored.

7.6.5 Sensitivity Analyses

Table 11. Sensitivity analyses – Inverse Probability of Censoring Weighting (IPCW)

Analysis Methods	This analysis will examine the impact of varying assumptions about the censoring-at-random condition on the estimated treatment effect. In the primary analysis, we assumed that censoring is independent of the outcome, conditional on the treatment, survival up to the time of censoring and indirectly baseline covariates used to estimate the treatment weights. Inverse probability of treatment weights were used to reweight the sample so that, in the weighted pseudo-population, treatment assignment is independent of measured covariates. In the IPCW analysis, inverse probability of censoring weights will be additionally applied to reweight observations. It assumes censoring is independent of the outcome, with all common causes of both the outcome and censoring being directly accounted for. Follow-up will be divided into equal 30-day intervals. At the start of each interval, we will update the information available on each patient and assess whether they remain followed or have been censored. If they remain under observation, they contribute to the next risk set for that interval. The weight for each participant at each interval is calculated as: 1/(the estimated probability of remaining uncensored (Which is estimated (as [1 - the conditional probability of being censored]) in the immediately previous interval), given a set of baseline and time-updated covariates that could affect both censoring and the outcome).

	Characteristics that could affect both censoring and the outcome include: treatment group (rivaroxaban, apixaban), demographics (age, sex), comedications (e.g., PPI use) and comorbidities (e.g., peptic ulcer disease) The weight is calculated separately for each interval, and then multiplied together across all intervals of follow-up to give each participant's cumulative weight. The denominator probability will be estimated using pooled logistic regression fit to the person-interval dataset. In this model, the outcome is whether the participant was censored in that interval. Time will be modelled flexibly using restricted cubic splines. We will truncate weights at prespecified percentiles (1st and 99th). The IPCW will be applied as time-varying weights in the Cox model for time to first major GI bleeding. Because inverse probability of treatment weights (IPTW) are also used, the final analysis weights will be the product of IPTW and IPCW.
Assumptions	- Censoring is **conditionally** independent of the outcome given covariates (i.e., non-informative censoring/censoring at random, conditional on other covariates beyond those in the analysis model). - Correct model specification - Positivity (there is a non-zero probability of remaining uncensored at each time point within each covariate pattern). - Outcome does not directly influence its own missingness (would imply informative censoring via MNAR mechanism).
What is Being Varied?	- The condition of the censoring at random assumption. - Tests a different missing-at-random (MAR) assumption
Why (Objective)	- To evaluate whether the treatment effect estimate is sensitive to changes in the condition of the missing at random assumption
Strengths	- IPCW adjusts for common causes (measured) of censoring and the outcome
Limitations	- IPCW is sensitive to model misspecification. - Cannot account for unmeasured factors affecting censoring. - Weighting can increase variance, especially if weights are unstable - Sensitivity analyses rely on varying the assumptions of the primary analysis, but for censoring these assumptions cannot be verified from the observed data; their plausibility can be discussed yet ultimately remains unknown.

Table 12. Sensitivity analyses – Best/Worst Case

<table>
<tr><td>Analysis Methods</td><td>The goal is to assess the impact on the estimated treatment effect of selected censoring non-at-random assumptions. The assumptions chosen represent the 4 extremes of a tipping point sensitivity analysis.

For non-administrative censored individuals, repeat the analysis under four scenarios, assuming within each treatment arm that a) all censored individuals had the outcome of interest at the censoring date (worst case), and b) none of the censored individuals had the outcome of interest by the end of the study (best case). This equates to the following scenarios:

<table>
<tr><th>Scenario</th><th>Exposed Group (Rivaroxaban)</th><th>Control Group (Apixaban)</th><th>Interpretation</th></tr>
<tr><td>1</td><td>Best case (lowest event rate)</td><td>Worst case (highest event rate)</td><td>Maximally favors Rivaroxaban</td></tr>
<tr><td>2</td><td>Worst case (highest event rate)</td><td>Best case (lowest event rate)</td><td>Maximally favors apixaban</td></tr>
<tr><td>3</td><td>Best case</td><td>Best case</td><td>Optimistic for both groups</td></tr>
<tr><td>4</td><td>Worst case</td><td>Worst case</td><td>Pessimistic for both groups</td></tr>
</table>
</td></tr>
<tr><td>Assumptions</td><td>• All or none of the censored individuals had the outcome of interest</td></tr>
<tr><td>What is Being Varied?</td><td>- In the primary analysis that censoring is non-informative given treatment group and survival up to the time of censoring.</td></tr>
<tr><td>Why (Objective)</td><td>- To assess the robustness of the treatment effect estimate to violations of the non-informative censoring assumption made in the primary analysis.
- If results are stable across scenarios, confidence increases that findings are not driven by bias (due to informative censoring).</td></tr>
<tr><td>Strengths</td><td>- Enables the exploration of alternative assumptions about censoring.
- Does not assume a certain structure of informative censoring.
- Sets bounds on the extent of maximum possible bias due to informative censoring</td></tr>
<tr><td>Limitations</td><td>- Best/worst case scenarios are extreme assumptions</td></tr>
</table>

Table 13. Sensitivity analyses – Non-Differential Exposure Misclassification

Analysis Method	***Probabilistic Bias Analysis using Monte Carlo Simulation at the summary estimate measure*** To evaluate the potential impact of exposure initiation misclassification at index date, we will conduct a probabilistic bias analysis at the summary estimate level. This analysis will be performed after pooling hazard ratios across imputed datasets using Rubin's rules so that uncertainty due to missing data and sampling variability is incorporated into the input effect estimate. The pooled hazard ratio and its variance will then serve as the starting point for the bias analysis. Plausible distributions for the sensitivity and specificity of prescription - based exposure classification will be specified based on the known limitations of prescription records as proxies for true medication use (e.g., prescriptions may not be filled or taken). This sensitivity analysis addresses misclassification of actual treatment initiation as inferred from prescription records (i.e. whether a recorded prescription corresponds to true initiation of use), and does not address misclassification of treatment discontinuation or adherence during follow-up. Within each Monte Carlo iteration, a hazard ratio will be sampled from a probability distribution informed by the pooled hazard ratio and its variance from the main analysis. This step propagates uncertainty due to sampling variability and multiple imputation into the bias analysis Final uncertainty will be represented by the percentile-based 95% confidence interval from the distribution of simulated bias-adjusted hazard ratios, which is derived from the empirical distribution of bias-adjusted point estimates across all Monte Carlo iterations and incorporates uncertainty from sampling variability, multiple imputation, and exposure misclassification, rather than from confidence intervals calculated within individual iterations. Plausible parameter ranges for prescription-based exposure measurement in real-world data (based on literature [12, 13]): Sensitivity: 0.70-0.90 o (lower bound reflects non-initiation/non-adherence after prescribing; upper bound reflects high adherence) Specificity: 0.95-0.99 o (high specificity expected, as those without prescriptions are unlikely to be truly exposed) There is no reason to suspect that misclassification of treatment initiation is differential with respect to the outcome, since prescribing occurs before the outcome; any association between prescribing and the outcome is more plausibly explained by common causes, which are addressed through IPTW. (in which prescribing would be independent of measured outcome determinants).

Assumptions	• Non-differential misclassification: exposure error is unrelated to the outcome. • Sensitivity and specificity are correctly specified and independent of the treatment arm. • The primary Cox regression model is correctly specified.
What is Being Varied?	• The assumption of accurate exposure measurement via prescribing records. • Simulate variation in sensitivity and specificity to generate corrected exposure measurement. • Assess how misclassification could bias the estimated hazard ratio.
Why (Objective)	• Evaluate whether the findings are robust to exposure misclassification. • Determine whether plausible levels of misclassification could meaningfully alter the estimated treatment effect.
Strengths	• Explicitly incorporates uncertainty in exposure classification. • Produces a distribution of corrected effect estimates, not just a point estimate.
Limitations	• Requires accurate or defensible assumptions about sensitivity and specificity. • Limited to non-differential misclassification unless extended. • Does not address misclassification of covariates or outcomes.

7.6.6 Other Supplemental Analyses

Baseline characteristics will be presented overall and stratified by treatment group. Categorical and binary variables will be summarized as counts (n) and percentages, while continuous variables will be reported using means and standard deviations or medians and interquartile ranges, as appropriate.

Kaplan–Meier methods will be used to compare the time-to-event distribution of GIB between patients treated with rivaroxaban and those treated with apixaban. Crude Kaplan–Meier cumulative incidence curves will be estimated separately for patients initiating rivaroxaban and for those initiating apixaban. The cumulative incidence (absolute risk) of major GIB will be estimated from the Kaplan–Meier curves at pre-specified time points of 1 and 2 years for each treatment group, together with 95% confidence intervals. Inverse probability of treatment weighted (IPTW) Kaplan–Meier curves will also be estimated. Time will be measured from the day after treatment initiation (index date) until the first occurrence of GIB or censoring.

The cumulative incidence (absolute risk) of GIB will be estimated using Kaplan–Meier methods over the entire follow-up period for each treatment group and presented graphically as full cumulative incidence curves. In addition, cumulative incidence will be reported at the pre-specified time points of 1 and 2 years, together with 95% confidence intervals. Values at 1 and 2 years will be obtained by evaluating the Kaplan–Meier step function at those time points

We will also conduct descriptive analyses to characterize censoring patterns overall and across treatment groups. This will include median (IQR) time to censoring overall and according to the reason for censoring. This will be estimated separately for the overall study population and by treatment arm.

Reasons for censoring will include:
Administrative censoring: reaching the maximum follow-up period of 2 years or the end of the study period.
End of data availability: last recorded healthcare encounter, database end date, or practice withdrawal.
Loss to follow-up: deregistration from the contributing practice or migration out of the healthcare system.

We will also compare baseline characteristics between eligible population and resulting study population after PS based trimming.

7.6.7 Core Emulation Tables – Estimation Summary

Table 14. Core Emulation Tables: Estimation Summary

Estimand 1

Attribute	Target Trial	Target Trial Emulation	Comment
Analysis Method	Cox proportional hazards model to estimate the hazard ratio for time to first major GI bleeding. Randomization ensures balance in measured and unmeasured confounders	Cox proportional hazards model weighted by stabilized IPTW, estimated separately in each data source (Danish registers and SIDIAP); pooled using random-effects meta-analysis. The analysis is conducted in in the trimmed population	IPTW used to emulate randomization in observational data. Trimming of observations based on PS distribution represents a departure from the original target trial but is considered best practice when using propensity score methods in emulation. By removing patients in regions of non-overlap, the analysis is restricted to a population where treatment assignment is more comparable across groups. As a result, the

			estimated effect no longer applies to the entire eligible population but to this more comparable subset. If some patients have an extremely low probability of receiving one of the treatments, valid causal contrasts cannot be identified for them. Without trimming, effect estimates in these regions rely on unsupported extrapolation, making the results unstable and potentially biased.
Missing Data Assumptions and Methods to Handle	**Outcome**: Assumes non-informative censoring conditional on treatment, and survival time; censored participants contribute partial information. **Exposure**: N/A (trial monitoring ensures exposure data completeness) **Covariates**: Minimized through trial data collection	**Outcome**: Same **Exposure:** For missing exposure data, assume absence of refill or prescription records for RIV or APX indicates true treatment discontinuation after 90 days. **Covariates:** absence of a diagnosis code will be interpreted as absence of the condition, while missing lifestyle and laboratory variables will be imputed using multiple imputation by chained equations (MICE) under the missing at random assumption.	Mechanisms of missing covariate and outcome data differ between the target trial and the emulation. In a target trial, follow-up and data collection are actively managed, whereas in the emulation, incomplete follow-up may arise due to practice deregistration, database end, or loss of data contribution (e.g., a GP practice no longer contributing data). Exposure data are assumed to be completely captured in the emulation based on prescription records and are not imputed. However, prescribing records may not perfectly reflect actual medication intake, which may introduce exposure misclassification. In a target trial, baseline covariates are actively collected, and missingness is typically minimized

			through protocolized data collection. In contrast, in the emulation, covariate data may be incomplete. Because these covariates are required for confounding adjustment, missing baseline covariate data will be addressed using multiple imputation..
Statistical Model Assumptions	Proportional hazards assumption for Cox model	Same; proportional hazards assumption assessed with Schoenfeld residuals and log(-log) plots. **IPTW Assumptions**: needed to identify causal effects: no unmeasured confounding, positivity, correct model specification, consistency	Diagnostics confirm appropriateness of Cox model; violations addressed in supplemental estimands and analyses (Restricted Mean Survival Time Analyses) Some assumptions for IPTW difficult to verify (e.g., unmeasured confounding). Can argue consistency may be violated as a result of allowing variables doses and medications as part of treatment arm.
Sensitivity Analyses	None	**IPCW:** Varies conditions of the censoring at random assumption(changed to include covariates (baseline and time-updated). **Best/worst case scenario analyses**, assuming CNAR **Probabilistic Bias Analysis:** Monte Carlo simulation to assess impact of non-differential exposure misclassification	Potential for exposure mismeasurement only present in emulation since the exposure measurement is based on prescription records, perfect assuming adherence to the prescribed treatment, which might not hold

Estimand 2

Attribute	Target Trial	Target Trial Emulation	Comment
Analysis Method	Accelerated failure time (AFT) model assuming Weibull distribution followed by estimation of the restricted mean survival time (RMST) at years 1 and 2	Weighted Weibull AFT model followed by estimation of the restricted mean survival time (RMST) at years 1 and2	
Missing Data Assumptions and Methods to Handle	Assumes non-informative censoring; censored participants contribute partial information under AFT	Same assumption; administrative and non-administrative censoring applied; multiple imputation for missing covariates	same assumptions about missing data are made and it is handled in the same way between target trial and emulation
Statistical Model Assumptions	Weibull survival distribution; log-linear relationship between covariates and log survival time	Same; assessed using diagnostics such as log(-log(S(t))) vs log(t) for Weibull assumption and Q-Q plot for residuals	AFT assumptions (e.g., distributional form, non-informative censoring) assessed graphically; cannot be verified empirically.
Sensitivity Analyses	None	None	NB: Sensitivity analyses (e.g., IPCW, tipping point, probabilistic bias analysis) only used for primary Cox model analysis for Estimand 1.

Estimand 3: Similar to estimand 1, but with non-administrative censoring of subjects discontinuing/switching treatment following a hypothetical approach, but without sensitivity analyses. IPCW (+IPTW) Cox models will be employed to allow such testing

7.7 Data sources

7.7.1 Data sources and quality

Rationale for selection and feasibility:

Denmark has a large network of population-based medical databases containing routinely collected data, covering many aspects of life and health. It covers all patients from birth to death, across all hospitals and medical clinics in the country. During decades, register data covering the total Danish population from cradle to grave have been collected. Most of this information has been collected for administrative purposes. Registers include redeemed prescriptions from community pharmacies, hospital admissions and outpatient contacts (diagnoses and procedures), laboratory measurements, migration data, socioeconomic data (income and education), cancer register, birth register and cause of death data [https://doi.org/10.2147/CLEP.S179083]

The Information System for the Development of Research in Primary Care (Sistema d'Informació per al Desenvolupament de la Investigació en Atenció Primària (SIDIAP) in Catalonia, Spain, is a primary care database set up by the Institute of Research in Primary Care (Institut Universitari D'Investigació en Atenció Primària Jordi Gol [IDIAP Jordi Gol]) and Catalan Institute of Health (Institut Català de la Salut). The database collects information from 328 primary health care centres and includes more than 5.8 million patients covered by the Catalan Institute of Health (approximately 78% of the Catalan population) and is highly representative of the Catalan population. SIDIAP data comprise the clinical and referral events registered by primary care health professionals (i.e., GPs, paediatricians, and nurses) and administrative staff in electronic medical records, comprehensive demographic information, community pharmacy invoicing data, specialist referrals, and primary care laboratory test results. SIDIAP can also be linked to other data sources, such as the hospital discharge database, on a project-by-project basis. Health professionals gather this information using International Classification of Diseases, 10th Revision (ICD-10) codes, ATC codes, and structured forms designed for the collection of variables relevant to primary care clinical management, such as country of origin, sex, age, height, weight, body mass index, tobacco and alcohol use, blood pressure measurements, and blood urine test results. In relation to vaccines, information on all routine childhood and adult immunisations is included in addition to the antigen and the number of administered doses.

Strengths of data sources:

Danish national registers:

The reliability of demographic data, hospital admission data, and overall diagnoses is deemed to be high as standard validation procedures are in place. Considering that the Danish population includes approximately 5.9 million inhabitants (as of 2023), the target sample size is anticipated to be reached.

SIDIAP:

All these practices use the same software, and all primary care health professionals receive similar training on the correct use of the software for optimal coding regarding clinical management of their patients Considering that SIDIAP includes data from approximately 5.8 million inhabitants, with 11,962 patients with non-valvular atrial fibrillation (NVAF) claimed a prescription of anticoagulation between 2011 and 2014 identified in previous literature, the target sample size is anticipated to be reached.

Limitations with potential impact in the study results:

Danish registers:

If the prescribed daily dose is not recorded, the defined daily dose (as defined by WHO) can be used as a proxy of consumption. Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.

SIDIAP:

Treatment duration and discontinuation needs to be estimated as this information is not readily avaiable in the data source.

Data quality:

Databases' suitability and case-study feasibility assessments followed three key steps: (I) characterization of data source systems and processes, using the EMA data quality checklist to evaluate foundational aspects and their maturity; (II) assessment of data quality metrics for each data source (data reliability), based on published research and open-access catalogues; and (III) fitness-for-use evaluation (data relevance), assessing database suitability for each case study based on question-specific determinants. Steps 1 and 2 were database-specific, while step 3 was both database- and case-specific, i.e. it could only be assessed in view of the specific research question to be addressed. From these steps, two tables containing qualitative information (I and III) and one with quantitative metrics (II) were created. The overall feasibility of the case studies using the candidate data sources was determined by critically analysing the collected information. Additional insights were gathered from DEAPs. All of the information was compiled into a report accompanying the generated tables, with our narrative assessment (appendix A).

The overall feasibility assessment is summarized in Table 15.

Table 15. Overall feasibility assessment summary for CS4 using Danish National Registers and SIDIAP

RWD source	Sample size estimation form the hypothetical trial protocol	Feasibility assessment *(yes/yes, with limitations/no)*	Rationale for the feasibility assessment	Limitations identified during the feasibility assessment and categorisation	Description of potential impact of the identified limitations on the study results
DNR	With an approximate samplze size required of 45,493 patients (i.e., 22,747 in each arm) and considering that the Danish population includes approximately 5.9 million inhabitants (as of 2023), the target sample size is anticipated to be reached.	Yes	Elements with high criticality are available and fairly reliable. Data recency is fine for extraction, reasonably enough for the research question. Sample size is achievable.	·Minor: Treatment duration and discontinuation needs to be estimated by means date of last medication acquisition. ·Minor: If the prescribed daily dose is not recorded, the defined daily dose (as defined by WHO) can be used as a proxy of consumption. ·Minor: Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.	As exact treatment duration is not available, depending on the method to estimate it we may under or overestimate exposure episodes.

SIDIAP	Considering that SIDIAP includes data from approximately 5.8 million inhabitants, with 11,962 patients with non-valvular atrial fibrillation (NVAF) claimed a prescription of anticoagulation between 2011 and 2014 identified in previous literature, the target sample size is anticipated to be reached.	Yes	Elements with high criticality are available and fairly reliable. Data recency is fine for extraction, reasonably enough for the research question. Sample size is achievable.	·Potentially major: Inability of capturing the cause of death in the database. Cause of death might need to be inferred since it is not recorded in the data source. ·Minor: Treatment duration and discontinuation needs to be estimated.	Since cause of death is not available, there is a risk of outcome misclassification. However, this limitation could be mitigated by inferring the likely cause of death based on diagnostic information recorded near the time of death. Additionally, because exact treatment duration is not available, estimates of exposure episodes may be under- or overestimated depending on the method used to approximate treatment duration.

Table 16. Metadata about data sources and software

	Data 1	Data 2
Data Source(s):	Danish registers	SIDIAP
Study Period:	1-1-2013-31-12-2023	1-1-2013-31-12-2023
Eligible Cohort Entry Period:	same	same
Data Version (or date of last update):		
Data sampling/extraction criteria:	Patients ≥75 years initiating rivaroxaban or apixaban	Patients ≥75 years initiating rivaroxaban or apixaban
Type(s) of data:	Data is collected from all hospitals and medical clinics in the country.	Key data: rare diseases, pregnancy and/or neonates, hospital admission and/or discharge, ICU admission, prescriptions of medicines, dispensing of reimbursed medicines, contraception, administration of vaccines, procedures, clinical measurements, healthcare provider, units of healthcare utilisation, unique identifier for persons, diagnostic codes, medicinal product information (active ingredient(s), dose, package size, strength), lifestyle factors (alcohol use, frequency of exercise, tobacco use), sociodemographic information (age, country of origin, deprivation index, gender, living in rural area, pharmaceutical copayment)
Data linkage:	Data can be linked to the country data set by ID. Also, to the Danish Clinical Quality Registries (RKKP), specific disease cohorts (cancer, depression, ADHD, surgery, cardiac arrest, among others). Additionally, whole families can be linked (mother-father-children).	Linkage with data augmentation with other data sources: - CMBD-URG (Hospital Emergency Room) - CMBD-AH (Hospital Discharges) - MHDA (Drugs Hospitalaries Dispensated in Ambulatory) - Pharmacies dispensations, EHR and Labaratories datasets
Conversion to CDM*:	TrineTX, ConcepTION, OMOP	ConcepTION, OMOP, NCDM (Nordic Common Data Model)
Software for data management:	No specific software	SQL/Python
HMA data catalogue link	https://catalogues.ema.europa.eu/institution/3331256	https://catalogues.ema.europa.eu/institution/50154

*CDM = Common Data Model

7.8 Data Management

The study will be conducted in a distributed manner using the UMCU, ARS Toscana and VAC4EU tools, procedures, and pipeline. Figure 1 specifies the data sets (D) and transformation processes (T), programming follows this pipeline, with involvement of different types of experts.

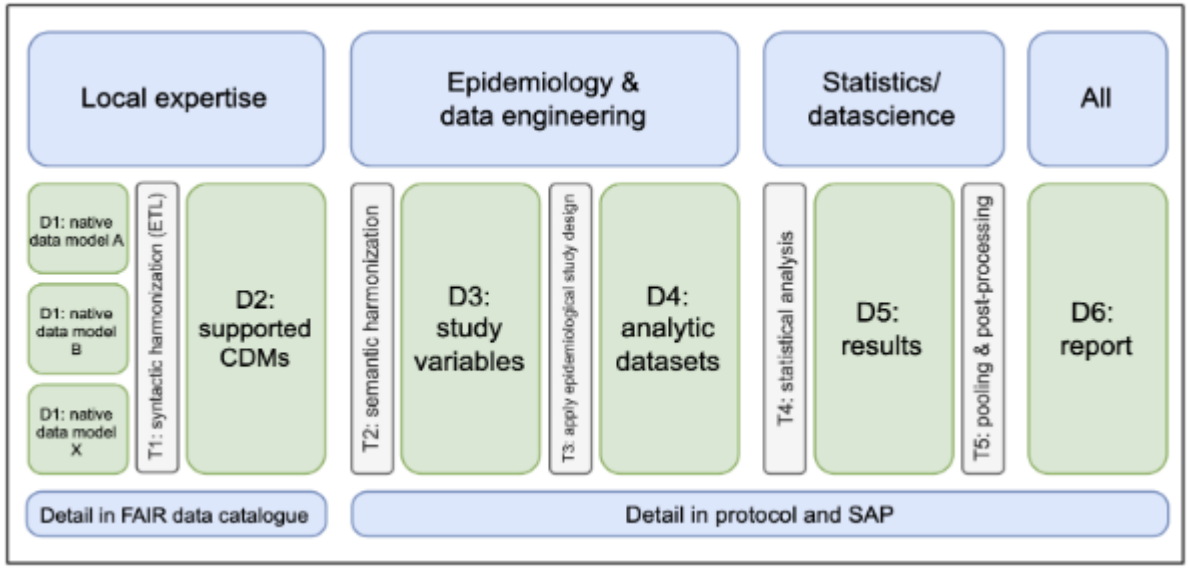


Figure 1. Data Management from the data transformation perspective

D1: Original data can be in any native format

The RWD-RWE pipeline used by VAC4EU starts with data banks that are controlled by the Data Expert and Access Partner (DEAP) and can be in any format. Data always stays local and never leaves the secure environments of the DEAPs. The ETL (extract, transform, load, see below for more details under 'T1') design is shared in a searchable FAIR VAC4EU catalogue. The VAC4EU FAIR Molgenis data catalogue is a meta-data management tool designed to contain searchable meta-data describing organisations that can provide access to specific data sources.

T1: Syntactic harmonisation (ETL)

T1: Syntactic harmonisation is conducted through an extraction, transformation, and loading (ETL) process of native data into the ConcePTION common data model (CDM) (see section 'D2: Common data model'). To harmonise the structure of the data sets stored and maintained by each data partner, a shared syntactic foundation is used. The ETL process has various structured steps as described by Thurin et al (2021):

- DEAPs are asked to share the data dictionaries of their data banks (selected tables and variable names/structure)
- Metadata (descriptive data about the data sources and databanks) & data dictionaries, are uploaded in FAIR data catalogue (Molgenis).

D2: Common data model

For this project, the CDM (D2) is the ConcePTION common data model. The CDM version that is used is v2.2, which is available as an open-source CDM. In this CDM, data are represented in a common structure, but the values of the data remain in their original language (e.g. codes will have either ICD9/10/ICPC/SNOMED or MEDCODEID values).

T2: Semantic harmonisation

During the T2 step, many data transformations occur related to the completion of missing features in the data. Based on the relevant diagnostic medical codes and keywords, as well as other relevant concepts (e.g., medications), one or more phenotype algorithms are constructed (typically one sensitive, or broad, algorithm and one specific, or narrow, algorithm) to operationalise the identification and measurement of each event. In this step we conduct time anchoring (observation periods, look back periods), clean the data such as the dose of vaccines, sort on record level, aggregate across multiple records, and combine concepts for implantation of algorithms, and rule-based creation of study variables.

In this phase of the creation of study variables, semantic mapping is conducted. This semantic mapping across different vocabularies is conducted as part of the R-study script using different functionalities. To reconcile differences between different terminologies and native data availability, machine-readable code lists are used that comprise the terminologies that are used in the network (e.g. ICD-9, ICD10, SNOMED, ICPC and DEAP specific adaptations). This is combined with the BRIDGE metadata file that defines risk windows, look-back periods, and algorithms for each study variable (65,66).

D3: Study variables

D3 datasets are interim data sets with information on study variables for each study participant, the unit may be a person, a medicine, or episode of time. The design of these datasets is described in codebooks. Examples of D3 datasets are the outputs of the ConcePTION pregnancy algorithm, and outputs of functions that define smoking. Multiple functions/packages exist within the VAC4EU, for different study variables.

T3: Application of epidemiological design

In the T3 step epidemiological designs are applied such as sampling, matching (on specific variables and/or propensity scores), and selection based on inclusion and exclusion criteria using the study variables in the D3 datasets. The designs will be implemented for the various study objectives using R-scripts, and these may use the existing functions (R-cran) or functions that have been developed in the VAC4EU community (e.g. matching).

D4: Analytical data set

D4 is an analytical dataset, and multiple D4 data sets may be produced based on the objectives of the study. The format is described initially in a code book for communication between programmers and statisticians.

T4: Statistical analysis

This step in the data transformation pipeline will produce statistical estimates such as descriptives (counts, percentages), distributions (mean, percentiles), rates (prevalence, incidence), regression coefficients, or other relevant estimates. This will be conducted using R.

D5: Results

D5 is the set of estimates, tables or aggregate data that is transferred from the DEAPs to the Digital Research Environment (DRE). The aggregated results produced by these scripts at the DEAP's site will be uploaded to the UMCU DRE for post-processing, pooling and visualisation (Figure 1). The DRE is a cloud-based, globally available research environment where data are stored and organised securely and where researchers can collaborate. The DRE is made available through UMCU. The DRE applies double authentication where researchers can collaborate using data that are stored and organised securely [ref]. UMCU is responsible for data processing and data security.

All researchers who need access to the DRE will be granted access to study-specific secure workspaces by UMCU. Access to the workspaces will be possible only after double authentication using an identification code and password together with the user's mobile phone for authentication.

Uploading files will be possible for all researchers with access to the workspace within the DRE. Downloading of files will be possible only after requesting and receiving permission from a workspace member with an “owner” role, who will be a UMCU team member.

T5: Post-processing/pooling

In this step, the result from different DEAPs is pooled and converted into tables and figures for reporting.

7.9 Quality Control

All key study documents such as the hypothetical trial protocol, target trial emulation protocol and study reports will undergo senior scientific and editorial review.

Data quality

For all data sources and for each data instance we will conduct *INSIGHT*[9] level 1-2 quality checks, detailed statistical analysis plans for the indicators are available on the public repositories:

- https://github.com/UMC-Utrecht-RWE/INSIGHT-Level1 Hoxhaj, V. (2023). UMC-Utrecht-RWE/INSIGHT-Level1: https://doi.org/10.5281/zenodo.10035167
- https://github.com/UMC-Utrecht-RWE/INSIGHT-Level2 Hoxhaj, V., & van den Bor, R. (2023). UMC-Utrecht-RWE/INSIGHT-Level2: https://doi.org/10.5281/zenodo.10035169

Briefly, level 1 verifies Data Completeness and level 2 Data Consistency.

Level 1 – Data Completeness

The purpose of the level 1 check is to verify the completeness of the ETL process and the data in the variables. Examples of tests are:

- Presence of variables in each of the CDM tables in D2
- Checks for misspellings and letter case in variable names in each of the CDM tables
- Verification of vocabularies
- Check date formats
- Check conventions of values

- Missing data analysis
- Frequency tables for categorical variables

Level 2 – Data Consistency

Real data is not random but follows certain logical constraints that reflect rules governing real-world situations. Examples of indicators generated by level 2 checks are:

- Event dates before date of birth
- Event dates after date of death
- Event dates out of observation periods
- Subjects having an observation but not present in the PERSONS table
- Observations associated with a visit id and occurred before/after the visit start/end date
- Subjects younger than 12 years old reported as parents
- Age at the observation period older than 115 y old Data

Code Quality

These coding practices define how the TARGET programming team collaborates to write clean, reliable, and reproducible code for the VAC4EU Real-World Evidence (RWE) Analytical Pipeline. They aim to ensure clarity, consistency, and maintainability across all case studies within the project.

Coding conventions

To ensure clarity, consistency, and maintainability across the project, the following conventions will be applied to all codebases within the project:

- Consistent style: Code follows a consistent and readable style (see the tidyverse style guide for R).
- Meaningful names: Use clear, descriptive names for variables, functions, and files to convey their purpose.
- Modular code: Break down code into small, reusable functions where possible.
- No hardcoded paths: Use configuration files or relative paths to ensure portability.

Following these conventions makes the code easier to understand, test, and reuse across case studies and teams.

Documenting Code

Code documentation is used to promote good coding practices and ensure our work is understandable, maintainable, and reproducible. To achieve this, we will:

- Use descriptive comments that explain the purpose and rationale behind code sections, focusing on why something is done, not just what.
- Clearly document function inputs, outputs, and side effects, using standardized formats (e.g., roxygen2 in R) where appropriate and supported.
- Write meaningful variable and function names to make the code as self-explanatory as possible.

Version Control

We use Git and GitHub to manage version control. These tools support good coding practices by enabling collaboration, tracking changes, accessing a project's history, and ensuring code quality through review and documentation.

A dedicated GitHub organisation has been created for the project (https://github.com/target-roc19). Each case study is managed in its own repository within this organisation. Repositories are structured consistently across case studies, to reinforce modularity. Access to repositories is controlled through teams.

During development, all repositories remain private to ensure confidentiality. Once the project is finalised, relevant repositories will be made public and assigned a digital object identifier (DOI) via Zenodo to support transparency, reproducibility, and reuse by the wider research community.

To maintain code quality and clarity, we follow the git and GitHub guidelines below.

- Always use pull requests (PRs): never push directly to the main branch.
- Open an issue before creating a new branch. Ideally, one PR resolves one issue to keep changes focused and reviewable.
- Every PR must be reviewed by at least one other person before merging.
- The PR author merges the PR after it has been reviewed and approved.
- Write clear, descriptive commit messages.
- Write informative PR descriptions, including:
    - A concise title
    - Links to related issues

- A summary of the changes

Continuous Integration

Continuous Integration (CI) is set up to automatically check code quality and run tests whenever changes are pushed to the repository or submitted through a pull request (PR). The CI workflow ensures that the package adheres to predefined style guidelines and that all automated tests pass before changes are merged.

Coding Template

Every case study follows the general coding template used across all code in the TARGET project. The folder structure is organised as follows:

case-study-template

|____data

| |____D2_cdm

| |____D3_study_variables

| |____D4_analytic_datasets

| |____D5_results

| |____D6_report

|____docs

|____logs

|____run

|____tests

|____transformations

| |____T2_semantic_harmonization

| |____T3_study_design

| |____T4_statistical_analysis

| |____T5_processing_results

|____CHANGELOG.md

|____LICENSE

|____README.md

Project Data Structure and Storage

The data folder follows the Real-World Evidence pipeline structure. Data conforming to the common data model is stored in the D2_cdm folder.

Results from transformations T2, T3, T4, and T5 are saved in the respective folders:

- D3_study_variables
- D4_analytic_datasets
- D5_results
- D6_report

Each dataset is associated with a codebook, explained in more detail below.

All data remain securely stored on the Data Expert and Access Partners (DEAPs) servers and are never transferred externally. For testing purposes, dummy datasets are created. These fall into two categories:

- Unit test data: Small, predefined input and output pairs used to test individual transformation steps. These are stored in the tests folder, not in data, and can support automated testing.
- Pipeline test data: Larger, more complex dummy datasets used to test whether the full pipeline runs as expected. These may be included in the repository only if they remain below GitHub's 100 MiB file size limit and will otherwise be shared via SharePoint.

Logging System

When the pipeline is executed, log files are saved in the logs folder. These logs are especially helpful when running the code in the DEAPs environment, as they help trace and diagnose potential errors. We recommend using the logger R package to handle logging throughout the pipeline. A sample logging setup can be found in the logger.R script located at the root of the project directory.

Executing the Analytical Pipeline

The run folder contains scripts used to execute each transformation step in the pipeline.

- A central script, run_pipeline.R, orchestrates the full pipeline from start to finish.
- Subscripts (e.g., run_T2.R or similar) are available to run individual transformation steps separately.

Typically, the run_pipeline.R script is the main entry point used by a DEAP to execute the full pipeline. Before running it in the DEAP environment, the pipeline may need to be adapted to local settings. This can be done using a configuration file that defines variables required to tailor the pipeline to a specific DEAP. Please note that configuration files should not include sensitive information.

Such a file might include variables like:

- The name of the DEAP
- The path to the local data instance
- The path to any required external resources

Testing and Quality Assurance

The tests folder contains scripts to test the analytical pipeline. Tests will be used to ensure code behaves as expected and remains stable over time. By systematically checking inputs, outputs, and edge cases, tests help catch errors early and make future changes safer. We use the testthat R package to structure and run unit tests.

Continuous integration (CI) is used to automate testing. With CI, tests are automatically run each time code is pushed to the repository (e.g., via GitHub Actions). This helps identify issues immediately, ensures that new changes do not break existing functionality, and supports better collaboration by enforcing consistent code quality across contributors.

Modular Data Transformation Workflow

The transformations folder follows the Real-World Evidence pipeline structure. It contains the source code for all transformation steps, which is typically written in R. Each subfolder corresponds to a specific step in the pipeline (e.g., T2_semantic_harmonization, T3_study_design, T4_statistical_analysis, T5_processing_results) and includes the relevant scripts and helper functions for that step.

During the T2 step, a database is usually created (e.g., using DuckDB). This database can be queried using SQL, and it is recommended that all SQL queries be saved as clearly named, standalone SQL script files to ensure readability and reusability.

The purpose of the transformations folder is to structure and modularise the processing logic, making it easier to maintain, test, and reuse across different case studies. By organising code by transformation step, teams can work in parallel, increasing efficiency.

Changelog

A changelog will be kept for all notable changes in the project. Changelogs help track the evolution of the project over time, making it easier for collaborators to understand what has changed between versions. We follow the structure and best practices outlined in Keep a Changelog.

Codebooks

Before developing code, codebooks are created to describe each dataset (D) within the pipeline. A codebook is a comprehensive document that outlines the structure, contents, and metadata of a dataset. It serves as a detailed reference guide for anyone working with the data and plays a crucial role in guiding the development of the analytical pipeline by clearly defining both the inputs and expected outputs.

All codebooks are summarized in a central index file, which provides a high-level overview of the pipeline's structure. For each codebook, the index file includes:

- A brief description of its purpose,
- A list of the scripts used to generate the corresponding dataset,
- A description of the input datasets and input parameters required.

The datasets D2, D3, D4, and D5 are typically subdivided into multiple smaller transformation steps, each detailed within their respective codebooks. These smaller transformation steps ensure that each part of the pipeline is clearly scoped and well-documented.

In addition to supporting development, codebooks help ensure quality control by making transformation logic transparent and verifiable, and they enhance reproducibility by documenting exactly how data is structured and used throughout the analytical pipeline.

Deployment

The analytical pipeline is delivered to DEAPs as a GitHub release, tagged with a version number. Versioning follows the format: vYYYYMMDD.XX, where the date indicates the release date and XX denotes the sub-version or revision number.

Any deployment issues can be reported via the GitHub repository using the issues feature, where the programming team responsible for the R code will collaborate with the local DEAP to resolve them as needed.

Reproducibility

It is recommended to locally use the renv R package to maintain the R version and version of packages for reproducibility purposes.

At this time, however, using renv reliably across different systems and environments remains challenging. For this reason, we currently recommend its use only in local development setups.

We are actively monitoring developments in the R ecosystem related to cross-platform reproducibility. As soon as a more stable and portable solution becomes available, we will revisit this guidance and promote broader adoption.

Licensing

The code will be made available under an open source license.

README Guidelines

Each case study repository includes a README that covers the following points:

- Project Overview: brief summary of the study goals and key research questions.
- Background: context and rationale for the study.
- Repository Structure: Outline of main folders and their contents.
- Data Overview: Description of data sources, formats, and data privacy considerations.
- How to Run: Instructions for running the pipeline and key scripts, plus where outputs are saved.
- Testing: How to run tests to verify code functionality.
- Contributing: Guidelines for code contributions and issue tracking.
- License: Information about the code license.
- Contact: Who to reach out to for help or questions.

7.10 Study precision

In this non-interventional study, no hypothesis test will be performed. The focus is on the precision of the estimated treatment effect. Assuming the study size in each RWD source will be similar to the sample size of the target trial, the precision is estimated as described below.

Estimation of the precision of the HR

To estimate the expected 95% CI for a HR from a Cox proportional hazards model, the standard error (SE) of the log(HR) is derived from the total number of events.

Assumptions- Equal allocation to treatment groups.
- Large-sample normal approximation for log(HR).
- Symmetric CI on the log scale.

The confidence interval (CI) width for the hazard ratio (HR) can be calculated using the following formula:

$CI\_width\_HR = exp(\hat{\beta} + 1.96 \times SE) - exp(\hat{\beta} - 1.96 \times SE)$

Where:
- $\hat{\beta}$ is the log hazard ratio (log(HR))
- SE is the standard error of the log(HR)
- 1.96 is the z-score for a 95% confidence interval

Calculation of 95% CI

1. Assume equal allocation:
Number of events per group: d1 = d2 = d / 2

2. Calculate SE of log(HR):
SE[log(HR)] = sqrt(1/d1 + 1/d2) = sqrt(2/d)

3. Construct the 95% CI on log scale:
log(HR) ± 1.96 × SE[log(HR)]

4. Convert back to HR scale:
CI_HR = exp(log(HR) ± margin)

<u>% Precision= (upper limit of the CI/ assumed HR) -1</u>
We calculated for three scenarios:

Scenario 1
Calculation based on 630 events, HR=1.25

SE = sqrt(2 / 630) = .0563

Lower CI = 1.1193
Upper CI = 1.3960

% precision = 1.3960/1.1193=25.7%

Scenario 2

Calculation under the assumption that overall event rate is 10% lower (567), HR=1.25

Lower CI = 1.1126
Upper CI= 1.4043
% precision = 26.2%

Scenario 3
Calculation based on 30% fewer events (441), HR=1.25
Lower CI = 1.0954
Upper CI = 1.4264

% precision = 30.2

Table 17. % precision under different events counts

Scenario	Number of Events	Hazard Ratio (HR)	Lower CI (HR)	Upper CI (HR)	CI precision
Scenario 1	630	1.25	1.1193	1.3960	25.7%
Scenario 2	1070	1.25	1.1126	1.4043	26.2%
Scenario 3	823	1.35	1.0954	1.4264	30.2%

8. Limitation of the methods

A key limitation of emulating a target trial using real-world data is that it may not be possible to fully reproduce all features of the target trial. These limitations are inherent to the target trial emulation approach.

Population, Eligibility, and Setting

Eligibility criteria were operationalised using diagnostic and prescription codes, while chronic conditions were measured over the entire observable history. These pragmatic definitions may lead to incomplete ascertainment of comorbidities and risk factors compared with trial screening procedures. As a result, some patients included in the emulation may not have met the target trial's eligibility thresholds, if their AF disease was not properly recorded, or if the phenotyping algorithm did not properly capture it.

Treatment Conditions and Exposure

Background therapy is not restricted, reflecting real-world practice, but also introducing heterogeneity in treatment regimens. These factors represent departures from the controlled dosing and treatment monitoring procedures of a randomised trial.

Exposure is defined based on prescription records, which do not capture whether patients actually filled or took their medications. As a result, some individuals categorized as "exposed" may not have received the intended treatment. This limitation is acknowledged as a source of non-differential exposure misclassification, which would bias effect estimates toward the null. A sensitivity analysis will be implemented to assess the impact of varying levels of sensitivity and specificity in exposure classification, based on published estimates of prescription adherence.

Treatment Assignment and Follow-up

Randomisation cannot be emulated. Instead, inverse probability of treatment weighting is used to approximate exchangeability. While IPTW balances measured baseline covariates, it cannot account for unmeasured confounding. Moreover, trimming to address propensity score distribution non-overlap restricts the analysis to a subset of patients, meaning the estimated effects may not generalise to the entire eligible population.

In the emulation, follow-up begins at the date of first prescription rather than at randomisation. This aligns the start of follow-up with treatment initiation but differs from the conditions of the target trial, where treatment allocation is randomised, and follow-up begins at randomisation. In our study, treatment allocation reflects real-world prescribing decisions based on patient characteristics rather than random assignment, which may introduce confounding. Furthermore, follow-up is determined by data availability and practice registration rather than scheduled study visits, which may result in imperfect measurement of loss to follow-up.

Outcomes and Intercurrent Events

Endpoints are identified using validated code lists in real-world data sources. Although these definitions are well established, they are not identical to adjudicated outcomes in clinical trials. Misclassification of GI bleeding, or cause of death remains possible, which may bias effect estimates.

Strategies for handling intercurrent events (treatment discontinuation, switching) are implemented using prescription records. While these reflect real-world effectiveness, the accuracy of operational definitions (e.g., gaps of >90 days to define discontinuation) is limited by prescribing and refill practices. Misclassification of discontinuation or switching may bias estimates of while-on-treatment effects.

Loss to follow-up is defined using practice deregistration or database end, which are proxies for true loss. Although unique patient identifiers mitigate risks of missed de-registration, there remains potential for misclassification (e.g., patients who stop attending their practice but do not formally deregister).

Analysis Methods and Statistical Assumptions

Unlike the target trial, which relied on randomisation to achieve balance of baseline participant characteristics, the emulation employs IPTW in Cox and AFT models. This approach requires unverifiable assumptions: no unmeasured confounding, correct model specification, and positivity. Departures from these assumptions could bias results.

The Cox proportional hazards model assumes that the relative treatment effect is constant over time, which may not be appropriate. This assumption is explicitly tested using Schoenfeld residuals and log-log survival plots. In addition, a supplemental analysis is conducted using an accelerated failure time (AFT) model with a Weibull distribution to estimate restricted mean survival time (RMST) at fixed time points (1 and 2 years).

Some subgroups of patients may be very unlikely to receive one of the treatments, resulting in limited overlap in covariate distributions between RIV and APX initiators; therefore the propensity score distributions are examined for evidence of positivity violations T

Missing Data and Censoring

Unlike the trial, where exposure and covariates were actively collected, the emulation must rely on real-world data. Missing exposure data may occur through incomplete prescribing records, while covariates (e.g., lifestyle factors) are imputed under a missing-at-random assumption. Misclassification

is also possible when absence of a diagnostic code is assumed to reflect absence of a condition. These differences from trial data collection procedures may affect validity.

A key limitation of this study is the reliance on assumptions regarding the censoring mechanism. In the primary Cox model, we assume that censoring is non-informative, meaning that the probability of being censored is independent of the outcome, conditional on treatment assignment, baseline covariates, and indirectly on covariates included in the IPTW model and survival up to the time of censoring. This assumption may not always hold in practice. For example, patients may leave the database or deregister from a practice due to worsening health, which could be directly related to their risk of experiencing the outcome. If this relationship is not fully captured by measured covariates, effect estimates could be biased.

To address this, we perform sensitivity analyses using inverse probability of censoring weights (IPCW), which relax the primary assumption by conditioning not only on baseline confounders but also on time-varying covariates that predict both censoring and the outcome. While IPCW provides a more flexible and potentially more realistic adjustment, it remains dependent on correct model specification and the availability of sufficient data to capture predictors of censoring. If important determinants of loss to follow-up are unmeasured or poorly recorded, residual informative censoring may persist. In addition, a tipping point sensitivity analysis is conducted to assess how strong the association between censoring and outcome would need to be to change the study's conclusions.

Data Source Heterogeneity:

Danish Registers and SIDIAP differ in population coverage, healthcare systems, coding practices, and linkage availability, which may introduce heterogeneity in effect estimates. Analyses will be performed separately within each data source using harmonized definitions under the Conception CDM framework. If pooled estimates are produced, heterogeneity will be assessed using meta-analytic methods (e.g., I^2 statistics).

9. Protection of human subjects

This is a non-interventional study using secondary data collection and does not pose any risks for individuals. Each data source research partner will apply for an independent ethics committee review according to local regulations. Data protection and privacy regulations will be observed in collecting, forwarding, processing, and storing data from study participants. Patient information This study involves data that exists in an anonymized structured format and contains no patient personal information. All parties will comply with all applicable laws, including laws regarding the implementation of organisational and technical measures to ensure the protection of patient personal data. Such measures will include omitting patient names or other directly identifiable data in any reports, publications, or other disclosures, except where required by applicable laws. Patient personal data will be stored at DAPs in encrypted electronic form and will be password protected to ensure that only authorised study staff have access. DAPs will implement appropriate technical and organisational

measures to ensure that personal data can be recovered in the event of a disaster. In the event of a potential personal data breach, DAPs shall be responsible for determining whether a personal data breach has in fact occurred and, if so, providing breach notifications as required by law.

Patient consent

As this study does not involve data subject to privacy laws according to applicable legal requirements, obtaining informed consent from individuals is not required.

10. Reporting of adverse events

For studies in which the research team uses only data from automated healthcare databases, according to the International Society for Pharmacoepidemiology Guidelines for GPP. “Aggregate analysis of database studies can identify an unexpected increase in risk associated with a particular exposure. Such studies may be reportable as study reports but typically do not require reporting of individual cases. Moreover, access to automated databases does not confer a special obligation to assess and/or report any individual events contained in the databases. Formal studies conducted using these databases should adhere to these guidelines.” For non-interventional study designs that are based on secondary use of data, such as studies based on
medical chart reviews or electronic health records, systematic reviews, or meta-analyses, reporting of adverse events/adverse drug reactions is not required. Reports of adverse events/adverse drug reactions should only be summarized in the study report, where applicable. According to the EMA Guideline on GVP, Module VI – Management and Reporting of Adverse Reactions to Medicinal Products, “All adverse events/reactions collected as part of [non-interventional post-authorization studies with a design based on secondary use of data], the submission of suspected adverse reactions in the form of [individual case safety reports] is not required. All adverse events/reactions collected for the study should be recorded and summarized in the interim safety analysis and in the final study report.” Module VIII – Post-Authorization Safety Studies echoes this approach. Legislation in the EU further states that for certain study designs such as retrospective cohort studies, particularly those involving electronic health records, it may not be feasible to make a causality assessment at the individual case level.

11. References

1. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham study. Stroke. 1991;22(8):983–988.

2. Giugliano RP. Non-vitamin K antagonist oral anticoagulants in older and frail patients with atrial fibrillation. Eur Heart J Suppl 2022;24(Suppl A):A1–A10.

3. Carnicelli AP et al. Direct Oral Anticoagulants versus Warfarin in Patients with Atrial Fibrillation: Patient-Level Network Meta-Analyses of Randomized Clinical Trials with Interaction Testing by Age and Sex. Circulation. 2022 Jan 5;145(4):242–255

4. Mamas MA, Batson S, Pollock KG, Grundy S, Matthew A, Chapman C, Manuel JA, Farooqui U, Mitchell SA. Meta-Analysis Comparing Apixaban Versus Rivaroxaban for Management of Patients With Nonvalvular Atrial Fibrillation. Am J Cardiol. 2022 Mar 1;166:58-64.

5. Yan, V.K.C., Li, HL., Wei, L. et al. Evolving Trends in Consumption of Direct Oral Anticoagulants in 65 Countries/Regions from 2008 to 2019. Drugs 83, 315–340 (2023)

6. Halperin JL, Hankey GJ, Wojdyla DM, et al. Efficacy and safety of rivaroxaban compared with warfarin among elderly patients with nonvalvular atrial fibrillation in the rivaroxaban once daily, oral, direct factor Xa inhibition compared with vitamin K antagonism for prevention of stroke and embolism trial in atrial fibrillation (ROCKET AF). Circulation 2014;130:138–46.

7. Hori M., Matsumoto M., Tanahashi N., Momomura S., Uchiyama S., Goto S., Izumi T., Koretsune Y., Kajikawa M., Kato M., Ueda H., Iwamoto K., Tajiri M., and on behalf of the J-ROCKET AF study investigators, Safety and efficacy of adjusted dose of rivaroxaban in Japanese patients with non-valvular atrial fibrillation, Circulation Journal. (2013) 77, no. 3, 632–638,

8. Halvorsen S, Atar D, Yang H, De Caterina R, Erol C, Garcia D, Granger CB, Hanna M, Held C, Husted S, Hylek EM, Jansky P, Lopes RD, Ruzyllo W, Thomas L, Wallentin L. Efficacy and safety of apixaban compared with warfarin according to age for stroke prevention in atrial fibrillation: observations from the ARISTOTLE trial. Eur Heart J. 2014 Jul 21;35(28):1864-72.

9. David A. Garcia, Deborah A. Fisher, Hillary Mulder, Lisa Wruck, Raffele De Caterina, Sigrun Halvorsen, Christopher B. Granger, Claes Held, Lars Wallentin, John H. Alexander, Renato D. Lopes, Gastrointestinal bleeding in patients with atrial fibrillation treated with Apixaban or warfarin: Insights from the Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation (ARISTOTLE) trial, American Heart Journal, Volume 221, 2020, 1-8

10. Gardarsdottir H, Souverein PC, Egberts TC, Heerdink ER. Construction of drug treatment episodes from drug-dispensing histories is influenced by the gap length. J Clin Epidemiol. 2010 Apr;63(4):422-7. doi: 10.1016/j.jclinepi.2009.07.001.

11. Schukman, S. et al. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. Journal of Thrombosis and Haemostasis 2005; Volume 3, Issue 4, 692 – 694

12. Charlton A, Vidal X, Sabaté M, Bailarín E, Martínez LML, Ibáñez L. Factors associated with primary nonadherence to newly initiated direct oral anticoagulants in patients with nonvalvular atrial fibrillation. J Manag Care Spec Pharm. 2021 Sep;27(9):1210-1220. doi: 10.18553/jmcp.2021.27.9.1210. PMID: 34464214; PMCID: PMC10391044.

13. Ozaki AF, Choi AS, Le QT, Ko DT, Han JK, Park SS, Jackevicius CA. Real-World Adherence and Persistence to Direct Oral Anticoagulants in Patients With Atrial Fibrillation: A Systematic Review and Meta-Analysis. Circ Cardiovasc Qual Outcomes. 2020 Mar;13(3):e005969. doi: 10.1161/CIRCOUTCOMES.119.005969. Epub 2020 Mar 9. PMID: 32148102.

14. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med. 2018 Jun 30;37(14):2252-2266. doi: 10.1002/sim.7654.