Master's Thesis: MSc Epidemiology

IMPLEMENTING THE COMPONENT ALGORITHM STRATEGY (CAS) USING ELECTRONIC HEALTH RECORDS IN A MULTI-DATABASE STUDY (MDBS)

Kanaka Vijay Soman

(Student Number: 4834720)

Under the supervision of:

Dr. Romin Pajouheshnia

Dr. Helga Gardarsottir

Dr. Patrick Souverein

Nicolas Hunt, PhD Student

Division of Pharmacoepidemiology and Clinical Pharmacology,

Utrecht University

Contents

Abstract:

Background: There is an increasing trend to use electronic healthcare databases to conduct drug safety studies since they are more efficient compared to primary data capture. EHR databases can still have less statistical power if the information on exposure and/or outcome of interest is rare to obtain a sufficient sample size. A way to solve this problem is by conducting a multi-database study (MDBS). MDBS pose several challenges including how to manage heterogeneity between the different included data sources. Therefore, it is necessary to find data source specific methodologies to conduct MDBS. One of the strategies is the Component Algorithm Strategy. Here, the choice of a particular event-finding algorithm is generally dependent on the characteristics of the data source. By creating different event-finding algorithms, we aimed to demonstrate the impact and usefulness of the component algorithm strategy by applying it to a case study which investigated the risk of GI bleeding and stroke in New Oral Anti-Coagulants (NOACs) versus Vitamin K antagonists (VKA) users.
Methods: The component algorithm strategy was developed using data from two European healthcare databases- CPRD Gold (UK) and PHARMO (The Netherlands). GI Bleeding and Stroke algorithms were created for available data domains (diagnoses, signs/symptoms), healthcare settings and concept sets (signs/symptoms for GI Bleeding and Non-traumatic, Traumatic and Unspecified for Stroke). Algorithm- and database-specific incident rates for study outcomes were estimated for the study period 2010 – 2019. Cox regression analysis was performed to calculate the risk of GI bleeding or stroke while using NOACs/VKAs and adjusted for confounders.
Results/Conclusion: The implementation of the component algorithm strategy had a visible impact on the cases retrieved by each algorithm, with sensitive algorithms detecting higher number of cases compared to more specific algorithms. When calculated for each event definition, the risk of GI bleeding and stroke was higher in DOAC users compared to VKA users in CPRD Gold data source, whereas there was an increased risk of GI bleeding or stroke in VKA users for some event definitions in the PHARMO database, whereas VKAs showed a protective effect for some other event definitions.

Plain Language Summary:

Although randomised controlled trials (RCTs) are the gold standard to assess the safety and effectiveness of a new pharmaceutical product, they exclude a wide variety of population and therefore it is necessary to assess the safety and effectiveness of the drugs also in the excluded populations. One of the solutions to solve this problem is making use of electronic health records (EHR) to and studying the utilisation and effectiveness of the drugs in the real world. However, these EHR databases pose some problems when the drug or the medical condition of interest is rare and there are not enough people to obtain a sufficient sample size. To overcome this, multiple data sources are used in a single study which has its strengths and limitations. One of the limitations is the differences between the two data sources known as database heterogeneity. This could be due to differences in the way data is collected within the two data sources. Therefore, it is necessary to develop strategies that are tailor made for each data source. One of the strategies, known as the Component Algorithm strategy, makes use of specific data domains present in the data source to identify events of interest. By creating such algorithms, we can aim to overcome the issue of database heterogeneity while retrieving outcomes of interest.

1. Introduction:

The safety and efficacy of drugs is proven through randomized clinical trials, but these trials often have limited power to detect rare or unknown adverse effects related to drug as the knowledge on the safety profile of drugs is incomplete at the time of market authorisation (*European Risk Management Strategy: 2008-2009 Work Programme Adopted*, 2007). More evidence needs to be provided by investigating the use of drugs in larger and more diverse populations by observing them for longer periods (Gini, Sturkenboom, et al., 2020). The EU Commission has implemented directives to increase the availability and quality of evidence on safety and effectiveness of drugs by mandating post-authorisation safety studies (PASS) throughout the life cycles of a drug (Pacurariu et al., 2018). There is an increasing trend to use electronic healthcare databases to conduct drug safety studies since they provide data on larger patient populations with longer follow-up periods and are also more efficient compared to primary data capture, i.e., collection of data specifically for a research study. (Miguel Coloma, n.d.)(Pacurariu et al., 2018).

Population-based electronic health records databases capture information on millions of patients. However, they can still have less statistical power if the information on exposure and/or outcome of interest is rare to obtain a sufficient sample size. A way to solve this problem is by combining several data sources in one study. (Gini, Sturkenboom, et al., 2020). To facilitate such studies, the European Medicines Agency (EMA) is in the process of setting up a coordination centre known as the Data Analysis and Real-World Interrogation Network (DARWIN-EU) that will enable the use of real-world healthcare databases across the European Union (EU). (ref. for EMA DARWIN-EU)(Reference for EMA DARWIN-EU, n.d.). Moreover, using multiple data sources in a single study helps to investigate the use and safety of drugs across countries with heterogenous populations, as well as aid in enhancing the generalizability of study results (Gini, Sturkenboom, et al., 2020). A multi-database study (MDBS) is conducted using multiple healthcare databases (at least two) in a single study. There is no overlap between the populations captured by the included data sources on an individual level, either because the databases are not linked to each other, or they come from different geographical regions and have different population characteristics (Gini, Sturkenboom, et al., 2020). MDBS thus includes various data sources such as primary care (data obtained from physicians, physician assistant, nurse, etc.), secondary care (eg: data hospital emergency department), death registries, insurance claims databases, research centres with multiple, linked data sources, etc., altogether providing a heterogenous mix of data. The use of different structures to record data, the coding systems or languages used can lead to differences in the scale of detail and meaning of the data captured by different data sources (eg: pharmacy dispensing vs GP prescribing)(Avillach et al., 2013).

To deal with database heterogeneity in a MDBS, it is necessary to develop strategies that are data source-specific to extract study population within a single source of electronic healthcare database (Roberto et al., 2016a). The first step in this strategy is to use a common definition to identify events of interest across all the databases, a process known as harmonization (Avillach et al., 2013). Based on this common or shared definition of event, a case finding algorithm (also known as event-finding algorithm) is then created, the choice for which depends on the characteristics of the data source in use as well as the research question of the drug safety study.

Since the primary aim behind data collection by electronic healthcare databases is providing patient care or administrative purposes, the quality of data may not be optimal when this data is used for research (primary aim for data collection ref.). The accuracy of a case-finding algorithm depends on the quality of information captured by the data source. (Bollaerts et al). While using routine clinical

care data for epidemiological research, there is a possibility that documentation of diagnosis or drug use is incorrect or incomplete. For instance, in majority of the databases, the hospital discharge files do not contain drug use information. Therefore, every hospital stay record is associated with missing drug exposure information by default and this information can be relevant for long periods of hospitalization (Schneeweiss & Avorn, 2005). Such misclassification of events leads to biases in epidemiological studies and affect the validity of the case-finding algorithms (Gini, Dodd, et al., 2020). To overcome this bias, one of the strategies is to retrieve events by combining information from different healthcare settings (eg: primary care and hospital) and different data domains available within a data source (eg: diagnosis, laboratory tests, or drugs for treatment of events). Combining information from different data domains in one algorithm will change the sensitivity and specificity of that algorithm compared to another which only uses information from one data domain. The choice of different event finding algorithms may help determine the accuracy of identifying events associated with each algorithm. In case of similar data domains across all the data sources, the event finding algorithms may also be able to tell if the differences in the outcome events are due to the differences in the underlying population captured by each data source.

Here, we investigated the possibility of constructing the component algorithm strategy by using a shared event definition and making use of the different data domains available within two selected databases from two European countries. The effect of the event-finding algorithms on the subsequent risk estimates of disease occurrence was compared across the databases. By estimating the number of events retrieved by each algorithm, we aim to provide information on the accuracy and usefulness of each algorithm.

2. Methods:

The component algorithm strategy was applied to a case study based on a published drug safety study commissioned by the EMA. The case study investigated the risk of major bleeding and stroke in users of novel oral anti-coagulants (NOACs)/direct oral anti-coagulants (DOACs) compared to users of Vitamin K antagonists (VKAs) with Non-Valvular Atrial Fibrillation (NVAF) in two European countries: United Kingdom and The Netherlands (Souverein et al., 2021a).

2.1 Data Sources:

This retrospective cohort study included data from two European healthcare databases: the Clinical Practice Research Datalink (CPRD) GOLD in the United Kingdom (UK) and the PHARMO Database Network in The Netherlands.

2.1.1. CPRD GOLD:

The Clinical Practice Research Datalink (CPRD) GOLD consists of routinely collected electronic health records from over 600 general practices throughout the UK. Since its establishment in 1987, it covers over 11 million patients, from which around 5 million are active patients that contribute data. In the UK, patients are registered with a general practitioner (GP) who is responsible for the primary health care of the patients as well as referrals to specialists. Therefore, CPRD collates information on prescription details, demographics, clinical events, tests, specialist referrals, immunisations, hospital admissions, and death. Diagnoses are recorded using the version 2 Read codes, whereas prescription details are recorded using a product name and British National Formulary (BNF) code, with dosage instructions and quantity (Herrett et al., 2015).

2.1.2. PHARMO Database Network:

The PHARMO database network consists of a network of healthcare providers such as general practitioners (GPs), clinical laboratories, in- and out-patient pharmacies, thereby forming a population-based network of electronic health record (HER) databases from the primary and secondary healthcare settings in the Netherlands. Since its establishment in 1990, it covers over 10 million residents of the Netherlands, from which about 7 million are active patients. The data sources in PHARMO are combined using patient level linkages. For instance, dispensing records from out-patient pharmacy database (or community pharmacies) are linked to hospital admission records on a patient level. The information available in the GP dataset includes diagnoses, symptoms, prescription for medications and orders for lab tests. In PHARMO, diagnoses and symptoms from the primary care are coded using the International Classification of Primary Care (ICPC) codes, whereas the hospital diagnoses and the secondary discharge diagnoses are mapped according to the World Health Organisation's (WHO) International Classification of Diseases (ICD)-9 and ICD-10 coding systems. The out-patient pharmacy databases capture information on drug dispensing from community pharmacies that includes Anatomical Therapeutic Classification (ATC) code for drugs, brand names, dosage, duration of drug dispensing. The in-patient pharmacy captures information on drug dispensing during hospitalization events (Kuiper et al., 2020).

2.2 Study Population:

The study population included subjects who initiated NOACs or VKAs, aged >= 18, with a confirmed diagnosis of NVAF (in both primary and secondary care data in PHARMO and primary care data only in CPRD GOLD) during the period of 1st January 2010 to 31st December 2019. The subjects were also required to have been registered in their respective databases for at least a year before the start of the study period. The cohort entry date (index date) was set to be the date of the first NOAC or VKA prescription (CPRD) or dispensing (PHARMO). New drug use was defined initiation of a NOAC or VKA during the study period and no use of these drugs for at least 12 months prior to the cohort entry date. To attribute the use of NOACs or VKAs to the diagnosis and treatment of NVAF, we only included patients with a NVAF diagnosis within 90 days before or after the index date. Subjects were followed until the outcome, loss to follow-up, death, move or end of study whichever occurred at the earliest. Patients were excluded if they had the event of interest in the 365 days prior to and on the cohort entry date, had a history of valvular atrial fibrillation prior to initiating a NOAC or VKA (Souverein et al., 2021a).

2.3 Exposure:

The case study compared the New Oral Anti-Coagulants (NOACs)/Direct Oral Anti-Coagulants (DOACs) to Vitamin K Antagonists (VKAs). For PHARMO, we assumed the duration of each dispensing to be 28 days. In CPRD, duration was calculated based on the information available on quantity, daily dose, and strength of the substance present in the drug. For patients where information was not available on one of these variables, duration of prescription was assumed to be 28 days. The treatment episodes were then created by using methods described elsewhere (Gardarsdottir et al., 2010). A maximum of 30-day gap between the theoretical end date of one prescription or dispensing and the start date of the next prescription or dispensing was allowed to happen. Treatment switch was not considered valid if a patient switched from one type of NOAC to another but was considered valid if a patient switched from NOAC to VKA or vice versa. A new treatment episode was constructed if the gap between two subsequent prescriptions or dispensing extended beyond 30 days.

2.4 Outcome Definition:

The primary outcome was occurrence of major bleeding and was defined as symptomatic bleeding in an organ or other critical area according to the International Society on Thrombosis and Homeostasis (ref. ISTH). For this study, major bleeding included only gastrointestinal (GI) bleeding and signs/symptoms related to GI bleeding such as melena (dark tarry stools), haematochezia (passage of bright red blood along with stools), and haematemesis (blood in vomit) (Mayo Clinic, n.d.). Gastrointestinal bleeding included bleeding from oesophagus, stomach, or the upper part of the small intestine, as well as bleeding from colon, rectum, or anus (Medical News Today_Sites of UGIB, n.d.) (Ghassemi & Jensen, 2013)

The secondary outcome was occurrence of any type of stroke: ischemic or haemorrhagic, transient ischemic attack (TIA), or unclassified stroke but excluded all-cause mortality. Ischemic stroke was defined as an event when the blood supply to any part of the brain is blocked or reduced, leading to oxygen and nutrients deprivation, and resulting in immediate death of brain cells (Mayo Clinic_Ischemic Stroke, n.d.) (Medline Plus_Hemorrhagic Stroke, n.d.). Haemorrhagic stroke was defined as bleeding inside the brain due to rupture of a blood vessel, resulting in immediate death of brain cells (Medline Plus_Hemorrhagic Stroke, n.d.). These clinical definitions of the primary and secondary outcome were used to create the standardized event definitions which were used in both the databases to retrieve events. The outcomes were identified from the primary care and hospital data in PHARMO using the ICPC and ICD9/ICD10 codes respectively. In CPRD Gold, the outcomes were identified from the primary care data using Read codes.

2.5 Potential Confounders:

Concomitant medications considered as potential confounders included those that increase risk of bleeding such as corticosteroids, selective serotonin reuptake inhibitors (SSRIs), non-steroidal anti-inflammatory drugs (NSAIDs) and anti-platelet drugs. Comorbidities considered to be risk factors for stroke were prior stroke/TIA, pulmonary embolism (PE), deep venous thrombosis (DVT), hypertension, congestive heart failure, diabetes, and other cardiovascular diseases (including coronary heart disease, myocardial infarction, angina pectoris, aortic plaque, and peripheral artery disease). Comorbidities considered as risk factors for gastrointestinal bleeding were prior bleeding, hypertension, stroke/TIA, DVT/PE, Alcoholism, any malignancy (excluding malignant neoplasm of the skin), gastrointestinal ulcer, thrombocytopenia, liver disease (hepatic impairment). The use of concomitant medications was measured for the period of 6 months prior to the index date. Similarly, comorbidities were defined as any observed record detected for a subject during the period of 6 months prior to the index date (Souverein et al., 2021a). The plan for addressing the confounders is discussed further in this report.

3. The Component Algorithm Strategy:

There are three standard parts to construct a component algorithm: the healthcare setting involved in data collection (either Primary Care or Hospital), the data domain involved, and the set of concepts that are used to define the event of interest which may be associated with one or more medical codes (eg: ICD, Read, ATC).

To identify events in a healthcare data source, one or more data domains can provide information. Combining data from one or all the domains can help in building different event finding algorithms with different accuracy to detect the events (Roberto et al., 2016b). For this study, two data domains were used: diagnosis and symptoms to identify cases of GI bleed and Stroke/TIA. The concept sets were then created for each of the chosen data domains.

A list of all the medical conditions (concepts) that could be possibly defined as GI Bleeding or Stroke/TIA was generated, along with their ICD9/ICD10, ICPC, and Read Codes. The concepts were then separated per the definitions of primary and secondary outcomes. For primary outcome, the concepts were categorised into sets according to the site/organ of bleeding. The codes for signs and symptoms related to GI bleeding were also retrieved and constructed into two separate concept sets (Table 1) (Mayo Clinic, n.d.). For the secondary outcome, the concepts were categorised into sets according to the type of stroke: non-traumatic stroke, traumatic stroke, and unspecified stroke/TIA. (Table 2). Traumatic stroke included brain haemorrhage following an injury. This concept was included in the study because prior studies have cited that use of oral anticoagulants after a traumatic brain injury leads to progression of disease of traumatic injury or leads to incidence of a new haemorrhagic/stroke event (Matsushima et al., 2021)(Albrecht et al., 2014). The component algorithms were then created based on the method described elsewhere (Gini, Dodd, et al., 2020) (Table 3).

Table1: The concept sets related to Primary Outcome: Gastrointestinal Bleeding:

Concept Set Label	Concept Set Description	Concept
Bleeding, GI	Bleeding from a specific site	Bleeding of alimentary canal (oesophagus to anus)
Bleeding, signs	Signs suggesting possibility of GI bleeding	Haematuria, Blood in Vomit, Blood in Stool
Bleeding, symptoms	Vague signs suggesting possibility of a bleed.	Bruising, Anemia

Table 2: The concept sets related to Secondary Outcome: Stroke/TIA:

Concept Set Label	Concept Set Description	Concept
Non-traumatic Stroke	Diagnosis of Stroke (Ischemic/Hemorrhagic) without the involvement of trauma/injury	Ischemic Stroke Hemorrhagic Stroke
Unspecified Stroke and TIA	Diagnosis of Non-Traumatic stroke and/or Diagnosis of Transient Cerebral Ischemia	Ischemic Stroke Hemorrhagic Stroke Transient Cerebral Attack
Traumatic Stroke	Diagnosis of stroke following trauma/injury which is expected to be directly exacerbated due to the use NOACs/VKAs	Ischemic Stroke Hemorrhagic Stroke

3.1 Creation and Selection of Component Algorithms:

For this study, two different healthcare settings were selected: Primary Care and Hospital. The concepts were associated with each of the healthcare setting by considering the availability of information on that component within a particular setting. (eg: since there was no Hospital data available in CPRD, none of the concepts were associated with this setting in). The resulting list of "component algorithms" served as "rules" to identify patients within a particular data domain (Roberto et al., 2016b). Individual component algorithms were combined logically into composite algorithms by making use of Boolean operators OR and AND (Gini, Dodd, et al., 2020) (Roberto et al.,

2016b). The following tables enlist the different component algorithms developed to detect events of GI bleeding and Stroke/TIA:

Table 3.1: Component Algorithms for the Primary and Secondary Outcome (PHARMO):

Name	Setting	Data Domain	Concept Set
PC Bleeding, GI	Primary Care Practice	Diagnosis	Bleeding, GI
Inpatient Bleeding, GI	Hospital	Diagnosis	Bleeding, GI
PC Bleeding, Signs	Primary Care Practice	Diagnosis	Bleeding, signs
Inpatient Bleeding, Signs	Hospital	Diagnosis	Bleeding, signs
PC Bleeding, Symptoms	Primary Care Practice	Diagnosis/Symptoms	Bleeding, symptoms
Inpatient Bleeding, Symptoms	Hospital	Diagnosis/Symptoms	Bleeding, symptoms
PC Non-traumatic Stroke	Primary Care Practice	Diagnosis	Non-traumatic Stroke
Inpatient Non-traumatic Stroke	Hospital	Diagnosis	Non-traumatic Stroke
PC Unspecified Stroke/TIA	Primary Care Practice	Diagnosis	Unspecified Stroke and TIA
Inpatient Unspecified Stroke/TIA	Hospital	Diagnosis	Unspecified Stroke and TIA
PC Traumatic Stroke	Primary Care Practice	Diagnosis	Traumatic Stroke
Inpatient Traumatic Stroke	Hospital	Diagnosis	Traumatic Stroke

Table 3.2: Component Algorithms for the Primary and Secondary Outcome (CPRD Gold):

Name	Setting	Data Domain	Concept Set
PC Bleeding, GI	Primary Care Practice	Diagnosis	Bleeding, GI
PC Bleeding, Signs	Primary Care Practice	Diagnosis	Bleeding, signs
PC Bleeding, Symptoms	Primary Care Practice	Diagnosis/Symptoms	Bleeding, symptoms
PC Non-traumatic Stroke	Primary Care Practice	Diagnosis	Non-traumatic Stroke
PC Traumatic Stroke	Primary Care Practice	Diagnosis	Unspecified Stroke and TIA
PC Unspecified Stroke/TIA	Primary Care Practice	Diagnosis	Traumatic Stroke

4. Data Analysis:

Baseline characteristics of the study population were estimated as means along with the standard deviations or as percentages as applicable. Based on the output of each component, crude incident rates (IRs) per 1000 person years were estimated, where the number of events retrieved by the respective component constituted the numerator and the follow-up person time of those patients constituted the denominator. To describe how the composite algorithms function, the number of events associated with these algorithms were summarised to see if there is an impact on the number of cases detected per algorithm. Further, Cox proportional hazards regression analysis was

performed to investigate the association between the use of NOACs or VKAs and the risk of study outcomes, expressed as hazard ratios along with 95% confidence intervals (95% CI). The statistical analyses were carried out using R software.

5. Results:

5.1 Study Population:

The total number of patients identified in the CPRD and PHARMO databases were 74679 and 22236 respectively. The proportions of female users for both the drug classes were similar in the two study populations. The mean age of prescription and dispensing was also comparable (slightly lower for NOAC users in the PHARMO database). The characteristics of the study populations are summarized in the following table:

Table 4: Baseline characteristics for NOAC and VKA users within CPRD and PHARMO:

	CPRD Gold (UK)		**PHARMO (NL)**	
	DOAC	**VKA**	**DOAC**	**VKA**
	(n = 35660)	**(n = 39019)**	**(n=11723)**	**(n=10513)**
Females (%)	16019 (44.9)	17210 (44.1)	4822 (41.1)	4872 (46.3)
Follow-up in years (mean (SD))	1.33 (0.72)	1.55 (0.66)	2.02 (1.94)	1.45 (0.85)
Age at index-date (mean (SD))	74.45 (10.86)	73.69 (10.42)	69.29 (11.58)	73.75 (11.59)
DVT_PE (%)	573 (1.6)	813 (2.1)	37 (0.3)	23 (0.2)
Hypertension (%)	21332 (59.8)	23932 (61.3)	846 (7.2)	373 (3.5)
GI_Ulcer (%)	2050 (5.7)	2139 (5.5)	81 (0.7)	43 (0.4)
Cancer (%)	13754 (38.6)	13388 (34.3)	344 (2.9)	244 (2.3)
Heart Failure (%)	1955 (5.5)	2394 (6.1)	217 (1.9)	271 (2.6)
Diabetes (%)	6471 (18.1)	6800 (17.4)	10 (0.1)	7 (0.1)
Renal Failure (%)	6500 (18.2)	7827 (20.1)	197 (1.7)	167 (1.6)
CVD (%)	7312 (20.5)	8784 (22.5)	504 (4.3)	333 (3.2)
Liver Disease (%)	31 (0.1)	23 (0.1)	70 (0.6)	47 (0.4)
Alcohol Abuse (%)	4074 (11.4)	3024 (7.8)	536 (4.6)	347 (3.3)
Thrombocytopenia (%)	163 (0.5)	174 (0.4)	11 (0.09)	15 (0.14)

5.2 Incident Rates estimated by component algorithms:

Table 5: Risk of Major Bleeding and Stroke per 1000 person-years per component algorithm within CPRD Gold and PHARMO:

	CPRD GOLD (UK)				**PHARMO (NL)**			
	Events in DOAC Users	**IR (/1000 PY)**	**Events in VKA users**	**IR (/1000 PY)**	**Events in DOAC Users**	**IR (/1000 PY)**	**Events in VKA users**	**IR (/1000 PY)**
Component Algorithms (N and IR per 1000 PY)								

PC Bleeding, GI	95	2.0	101	1.6	39	6.8	15	9.2
PC Bleeding, Signs	94	1.9	114	1.8	3	0.5	2	0.9
PC Bleeding, Symptoms	0	0	0	0	35	6.2	12	8
PC Non-Traumatic Stroke	420	9.6	568	14.9	27	4.8	20	8.1
PC Traumatic Stroke	3	0.1	14	0.2	0	NA	NA	NA
PC Unspecified Stroke/TIA	15	0.3	13	0.2	31	5.5	33	10.9
Inpatient Bleeding, GI	NA	NA	NA	NA	43	1.8	36	2.8
Inpatient Bleeding, Signs	NA	NA	NA	NA	23	2.4	34	2
Inpatient Bleeding, Symptoms	NA	NA	NA	NA	3	0.1	9	0.4
Inpatient Non-Traumatic Stroke	NA	NA	NA	NA	0	0	6	0.2
Inpatient Traumatic Stroke	NA	NA	NA	NA	0	0	0	0
Inpatient Unspecified Stroke/TIA	NA	NA	NA	NA	6	0.3	10	0.5

The above table summarizes the incident rates associated with each component algorithm for both GI bleeding and Stroke in CPRD and PHARMO data sources. There were no diagnostic codes identified for the traumatic stroke conditions in the ICPC coding system, hence this component algorithm was not created for the PHARMO data source.

5.3 Number of cases (incidences) retrieved by composite algorithms per data source:

Table 6: Number of events retrieved by the composite algorithms in PHARMO:

Database	**PHARMO (NL)**			
Component Algorithms not stratified for drug class (N and IR per 1000 Person-Years)				
PC Bleeding, GI	61			
Inpatient Bleeding, GI	88			
PC Bleeding, Signs	6			
Inpatient Bleeding, Signs	61			
PC Bleeding, Symptoms	52			
Inpatient Bleeding, Symptoms	13			
PC Non-traumatic Stroke	54			
Inpatient Non-traumatic Stroke	7			
PC Unspecified Stroke/TIA	78			
Inpatient Unspecified Stroke/TIA	20			
PC Traumatic Stroke	NA			
Inpatient Traumatic Stroke	0			
Composite Algorithms (N)				

	N (IR)	N (IR) in LHS component	N (IR) in both components	N (IR) in RHS component
PC OR Inpatient Bleeding, GI	145	61		88
PC AND Inpatient Bleeding, GI	4	61		88
PC Bleeding, signs OR PC Bleeding, symptoms	57	6		52
PC Bleeding, signs AND PC Bleeding, symptoms	1	6		52
Inpatient Bleeding, signs OR Inpatient Bleeding, symptoms	74	61	0	13
Inpatient Bleeding, signs AND Inpatient Bleeding, symptoms	0	61		13
PC Bleeding, GI OR PC Bleeding, symptoms	110	61		52
PC Bleeding, GI AND PC Bleeding, symptoms	3	61		52
Inpatient Bleeding, GI OR Inpatient Bleeding symptoms	101	88	0	13
Inpatient Bleeding, GI AND Inpatient Bleeding symptoms	0	88		13
Inpatient Bleeding, GI OR Inpatient Bleeding, Signs	149	88	0	61
Inpatient Bleeding, GI AND Inaptient Bleeding, Signs	0	88		0
PC Bleeding, GI OR PC Bleeding, Signs	67	61	0	6
PC Bleeding, GI AND PC Bleeding, Signs	0	61		6
PC Bleeding, GI OR PC Bleeding, Signs OR PC Bleeding, Symptoms	115	61		57
PC Bleeding, GI AND PC Bleeding, Signs OR PC Bleeding, Symptoms	3	61		57
PC Bleeding, GI AND Bleeding, signs AND Bleeding, symptoms	0	88	0	1
Inpatient Bleeding, GI OR Bleeding, Signs OR Bleeding, Symptoms	162	88	0	74
Inpatient Bleeding, GI & Bleeding, Signs OR Bleeding, Symptoms	0	88		74
Inpatient Bleeding, GI AND Bleeding, Signs AND Bleeding, symptoms	0	61		0

PC Bleeding, Signs OR PC Bleeding, Symptoms OR Inpatient Bleeding, Signs OR Inpatient Bleeding, Symptoms	131	57	0	74
PC Bleeding, GI OR (Any Setting Bleeding Signs OR Any settings Bleeding, Symptoms)	189	61		131
PC Bleeding, GI AND (Any Setting Bleeding Signs OR Any settings Bleeding, Symptoms)	3	61		131
Inpatient Bleeding, GI OR (Any Setting Bleeding Signs OR Any Bleeding Symptoms)	215	88		131
Inpatient Bleeding, GI AND (Any Setting Bleeding Signs OR Any Bleeding Symptoms)	4	88		131
PC Nontraumatic Stroke OR Inpatient Non-Traumatic Stroke	61	54		7
PC Non-Traumatic AND Inpatient Non-Traumatic Stroke	0	54		7
PC OR Inpatient Unspecified Stroke/TIA	97	78		20
PC AND Inpatient Unspecified Stroke/TIA	1	78		20
PC Non-Traumatic OR Unspecified Stroke/TIA	131	54		78
PC Non-Traumatic AND Unspecified Stroke/TIA	1	54		78
Inpatient Non-Trauamtic OR Unspecified Stroke/TIA	27	7	0	20
Inpatient Non-Trauamtic AND Unspecified Stroke/TIA	0	0	0	0

The above table summarises the number of events identified and the associated IRs for all the composite algorithms generated in the PHARMO database. The number of cases retrieved by the component “PC Bleeding, GI” were 61, whereas the number of cases retrieved by the component “Inpatient Bleeding, GI” were 88. Combining these two components into the composite “PC OR Inpatient Bleeding, GI” increased the number of cases retrieved to 145. There were 4 cases of GI Bleeding included in both the healthcare settings. The composite “PC Bleeding Signs OR Bleeding, Symptoms” identified 57 cases. On combining these two components with “PC Bleeding, GI” with the OR operator, the number of cases that were identified increased to 115. Also, 3 cases were detected when “PC Bleeding, GI” was combined using “AND” operator with the composite “PC Bleeding Signs OR Bleeding, Symptoms”. The number of cases detected by the algorithm “PC Bleeding, Signs OR PC Bleeding, symptoms” were 57. For the hospital setting, the composite “Inpatient Bleeding, Signs OR Inpatient Bleeding, Symptoms” detected 74 cases. Combining the signs and symptoms of bleeding from both the healthcare settings resulted in higher number of cases detected (n= 131). Further combining this information with cases of bleeding recorded with a diagnostic code resulted in detection of 189 cases in the PC setting and 215 cases in the hospital setting. The idea that signs or symptoms can be captured in either of the settings and the actual diagnosis could be in one of the two settings was observed through this algorithm. Although the

number of cases retrieved are less, it could be interesting to combine signs and symptoms information with the outcome of interest information in future studies. For the secondary outcome Stroke/TIA, similar results were obtained where more sensitive algorithm detected more cases compared to a stricter or more specific algorithm. There was no component created for signs and symptoms related to Stroke as these conditions are quite broad and for the sake of this study, it was quite out of scope to associate them with the stroke outcomes. While analysing the results for the composite algorithms, it was observed that a few algorithms did not retrieve any events such as the composite "PC Bleeding, Signs AND PC Bleeding, Symptoms" or "PC Non-Traumatic Stroke AND PC Traumatic Stroke" for CPRD Gold and "Inpatient Bleeding, Signs AND Inpatient Bleeding, Symptoms" for PHARMO. This is consistent with the expectation that stricter event definitions tend to identify a smaller number of cases. Since these algorithms were combined using the logical connector AND, they required events to be identified by both the components of the composite, making the event definition too specific.

Table 7: Number of events retrieved by the composite algorithms in CPRD Gold:

	CPRD GOLD (UK)			
Component Algorithm (N)				
PC Bleeding, GI	202			
PC Bleeding, Signs	216			
PC Bleeding, Symptoms	0			
PC Non-Traumatic Stroke	1027			
PC Traumatic Stroke	18			
PC Unspecified Stroke/TIA	32			
Composite Algorithms (N)				
	N for the Composite Output	**N in LHS Component**	**N in Both Components**	**N in RHS Component**
PC Bleeding, GI OR PC Bleeding, Signs	405	202	13	216
PC Bleeding, GI AND PC Bleeding, Signs	13	202		216
PC Traumatic Stroke OR Unspecified Stroke/TIA	49	18	1	32
PC Traumatic Stroke AND Unspecified Stroke/TIA	1	18		31

The above table summarises the number of events identified and the associated IRs for all the composite algorithms generated in the CPRD Gold database. A similar trend was observed with the CPRD data, where the composite algorithms created using the OR operator retrieved more cases, while the more specific algorithms created using the AND operator retrieved only those cases that were common both the components of the composite algorithm.

5.4 Association between the risk of GI Bleeding or Stroke and the use of NOACs or VKAs:

Table 8: Risk of GI Bleeding and Stroke in VKA users compared to NOAC users expressed as Hazard Ratios:

Algorithm	**Hazard Ratio (95% CI)**
PHARMO	
Inpatient Bleeding, Signs	2.0 (1.15 - 3.5)
PC Bleeding, Signs OR PC Bleeding, Symptoms	0.52 (0.28 - 0.97)
Inpatient Bleeding, Signs OR Inpatient Bleeding, Symptoms	1.88 (1.16 - 3.06)
PC Bleeding, GI OR PC Bleeding, Symptoms	0.56 (0.35 - 0.88)
PC Bleeding, GI OR PC Bleeding, Signs OR PC Bleeding, Symptoms	0.59 (0.38 - 0.92)
Inpatient Bleeding, GI OR Inpatient Bleeding, Signs OR Inpatient Bleeding, Symptoms	1.47 (1.05 - 2.04)
PC Unspecified Stroke/TIA	1.87 (1.10 - 3.2)
CPRD Gold	
PC Bleeding, GI OR PC Bleeding, Signs OR PC Bleeding, Symptoms	0.78 (0.64 - 0.96)
PC Bleeding, GI OR PC Bleeding, Symptoms	0.75 (0.57 -0.98)
PC Bleeding, GI OR PC Bleeding, Signs	0.78 (0.64 - 0.96)
PC Non-Traumatic OR PC Traumatic Stroke	0.79 (0.70 - 0.90)
PC Non-Traumatic OR PC Unspecified Stroke/TIA	0.77 (0.68 - 0.87)

We estimated the association between the use of VKAs and the risk of bleeding or stroke compared to the use of NOACs for each event definition using Cox proportional hazards regression models adjusted for comorbidities and gender. Overall, in CPRD there was a statistically significant decreased risk of GI bleeding and stroke in VKA users as compared to NOAC users. By contrast, in PHARMO, some algorithms reflected a statistically significant increased risk of GI bleeding and Stroke with the use of VKAs compared to the NOACs, while some algorithms showed a decreased risk of study outcomes while using VKAs versus NOACs.

6. Discussion:

This multi-database study aimed at designing and implementing different algorithms to detect the cases of GI bleeding and stroke and estimate the IRs in two European healthcare databases. The number of cases detected by these algorithms showed heterogeneity within as well as between the two data sources. The algorithms designed to detect the cases of GI bleeding consistently estimated lower IRs compared to the published literature (Souverein et al., 2021b) in both the databases. This could be due to a very strict overall definition of the outcome GI bleeding, which excluded all the types of bleeding that were not a part of the alimentary canal, as well as the conditions (signs or symptoms) not related to or suggesting a possible GI bleeding. For the secondary outcome, the IRs in CPRD and PHARMO were comparable to that reported in published literature (Souverein et al., 2021b).

Further, broader algorithms that were not too specific retrieved a higher number of cases compared to the more specific ones. However, the specific algorithms were useful to identify the events that were common to more than one healthcare setting or data domain (Gini, Dodd, et al., 2020). Therefore, it might be interesting to take this into consideration while calculating the effect estimates as it might overestimate the results by counting the same case more than once. On the other hand, combining the signs and symptoms information from both PC and Hospital settings resulted in a higher number of cases detected as compared to using only one healthcare setting.

The different coding systems used to capture the diagnoses and signs/symptoms information in CPRD and PHARMO might have also affected the incidence rates of the outcomes of interest in this study. For instance, there were no events retrieved by the event definition "PC Bleeding, Symptoms" in CPRD Gold, which uses the Read coding system. On the other hand, symptoms information was captured to some extent in the PHARMO data source that uses the ICPC coding system.

As the aim of this methodological study was to investigate the use of different components to detect outcomes, and not to provide the best assessment of the comparative safety and effectiveness of NOACs, time-varying confounding was not addressed during the analysis. The purpose behind this was to simplify the interpretation of the analysis results. Thus, time-invariant confounding was accounted for during the analysis and adjusted for in the Cox proportional hazards regression analysis. There is no reason to suggest that not adjusting for time-varying confounders would have influenced the results of the comparison of different event-finding algorithms.

7. Strengths and Limitations:

The strength of this study was the implementation of the component algorithm strategy simultaneously to two data sources combined in a study by using a common event definition for the primary and secondary outcomes. Exploring different components provided an insight into the granularity of data collection within each of the involved data sources and could be a potential tool to address heterogeneity encountered while conducting multi-database pharmacoepidemiologic studies. The study investigated the exposure to NOACs or VKAs per each treatment episode and investigated the occurrence of study outcomes within the treatment episode. Therefore, the outcomes observed in this study can be fully attributed to the exposure to either NOACs or VKAs.

Further, it was possible to investigate the association between risk of GI bleeding or stroke and the use of NOACs or VKAs separately in two different countries in Europe. The difference in the hazard ratios between the two study populations potentially highlights the very reasons of conducting a multi-database study: investigating the differences underlying the two populations, the differences in the policies or patterns of drug use in these countries, etc. The results obtained in PHARMO showing increased risk for GI bleeding or stroke among VKA users for some event definitions compared to others warrants further investigation, but also highlights the possibility that these differences exist due to the differences in the way algorithms retrieved events in these study populations.

There are also a few limitations to this study. As there was no HES data from CPRD available, the results from the hospital setting of PHARMO could not be compared to a similar level. Due to time constraints, the hazard ratios were not fully adjusted to account for confounding due to concomitant medication use. Further, there could be some exposure misclassification since the two data sources utilised different methods to estimate exposure duration. Also, drug dispensing information during hospitalisations was not available for analysis, thus not accounting for the exposure during hospitalisation. The differences in the coding systems employed by the two data sources could have

resulted in some outcome misclassification. Lastly, the differences in the healthcare systems of the UK and the Netherlands reflecting in the differences in the drug prescribing policies as well as the utilisation of healthcare could have impacted the comparison between the two data sources involved in this study.

8. Scope of the Component Algorithm Strategy:

The component algorithm strategy has a wider scope that goes beyond this study. There is potential to apply this strategy by incorporating more data domains such as the diagnostic tests or laboratory results or drug treatments used for managing GI bleeding and stroke and combining these with the signs or symptoms data domains to identify any cases that were not recorded as a diagnosis accompanied by a diagnostic code. The algorithms can be further validated by estimating the positive predictive value (PPV) associated with each algorithm using the true proportion of cases and the proportion of cases detected by the algorithm itself (Gini, Dodd, et al., 2020). In case of a missing data domain in one database, it might be interesting to see if it is possible to predict outcomes by using another database that has information on this missing data domain.

References:

1. Albrecht, J. S., Liu, X., Baumgarten, M., Langenberg, P., Rattinger, G. B., Smith, G. S., Gambert, S. R., Gottlieb, S. S., & Zuckerman, I. H. (2014). Benefits and risks of anticoagulation resumption following traumatic brain injury. *JAMA Internal Medicine*, *174*(8), 1244–1251. https://doi.org/10.1001/jamainternmed.2014.2534
2. Avillach, P., Coloma, P. M., Gini, R., Schuemie, M., Mougin, F., Dufour, J. C., Mazzaglia, G., Giaquinto, C., Fornari, C., Herings, R., Molokhia, M., Pedersen, L., Fourrier-Réglat, A., Fieschi, M., Sturkenboom, M., van der Lei, J., Pariente, A., & Trifirò, G. (2013). Harmonization process for the identification of medical events in eight European healthcare databases: The experience from the EU-ADR project. *Journal of the American Medical Informatics Association*, *20*(1), 184–192. https://doi.org/10.1136/amiajnl-2012-000933
3. *European Risk Management Strategy: 2008-2009 work programme adopted*. (2007). http://www.emea.europa.eu
4. Gardarsdottir, H., Souverein, P. C., Egberts, T. C. G., & Heerdink, E. R. (2010). Construction of drug treatment episodes from drug-dispensing histories is influenced by the gap length. *Journal of Clinical Epidemiology*, *63*(4), 422–427. https://doi.org/10.1016/j.jclinepi.2009.07.001
5. Ghassemi, K. A., & Jensen, D. M. (2013). Lower GI bleeding: Epidemiology and management topical collection on large intestine. *Current Gastroenterology Reports*, *15*(7). https://doi.org/10.1007/s11894-013-0333-5
6. Gini, R., Dodd, C. N., Bollaerts, K., Bartolini, C., Roberto, G., Huerta-Alvarez, C., Martín-Merino, E., Duarte-Salles, T., Picelli, G., Tramontan, L., Danieli, G., Correa, A., McGee, C., Becker, B. F. H., Switzer, C., Gandhi-Banga, S., Bauwens, J., van der Maas, N. A. T., Spiteri, G., … Sturkenboom, M. (2020). Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project. *Vaccine*, *38*, B56–B64. https://doi.org/10.1016/j.vaccine.2019.07.045
7. Gini, R., Sturkenboom, M. C. J., Sultana, J., Cave, A., Landi, A., Pacurariu, A., Roberto, G., Schink, T., Candore, G., Slattery, J., & Trifirò, G. (2020). Different Strategies to Execute Multi-Database Studies for Medicines Surveillance in Real-World Setting: A Reflection on the European Model. In *Clinical Pharmacology and Therapeutics* (Vol. 108, Issue 2, pp. 228–235). Nature Publishing Group. https://doi.org/10.1002/cpt.1833

8. Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., Staa, T. van, & Smeeth, L. (2015). Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology*, *44*(3), 827–836. https://doi.org/10.1093/ije/dyv098
9. Kuiper, J. G., Bakker, M., Penning-Van Beest, F. J. A., & Herings, R. M. C. (2020). Existing data sources for clinical epidemiology: The pharmo database network. *Clinical Epidemiology*, *12*, 412–422. https://doi.org/10.2147/CLEP.S247575
10. Matsushima, K., Leichtle, S. W., Wild, J., Young, K., Chang, G., Demetriades, D., Lomangino, C. D., Massetti, J., Glass, N. E., Livingston, D. H., O'Bosky, K., Chan, J. L., Cullinane, D. C., Rippey, K. A., Sharma, J., Nahmias, J., Grigorian, A., Allen, S., Armen, S. B., … Sabour, A. F. (2021). Anticoagulation therapy in patients with traumatic brain injury: An Eastern Association for the Surgery of Trauma multicenter prospective study. *Surgery (United States)*, *169*(2), 470–476. https://doi.org/10.1016/j.surg.2020.07.040
11. Mayo Clinic. (n.d.). *Signs & Symptoms_GI Bleed*. Retrieved March 18, 2022, from https://www.mayoclinic.org/diseases-conditions/gastrointestinal-bleeding/symptoms-causes/syc-20372729
12. Mayo Clinic_Ischemic Stroke. (n.d.). *Mayo Clinic_Ischemic Stroke*. Retrieved March 28, 2022, from https://www.google.com/search?q=ischemic+stroke&rlz=1C1CHZN_enIN982IN982&sxsrf=APq-WBtEHiH_WFhgLUQeAzdsQ17T8_38VQ%3A1648456258513&ei=QnJBYo3uHs-tkwXx1bXQCw&oq=ischemi&gs_lcp=Cgdnd3Mtd2l6EAEYADIECAAQQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIFCAAQgAQyBQgAEIAEMgUIABCABDoECCMQJzoLCC4QgAQQxwEQowJKBAhBGABKBAhGGABQAFiNBmDFEWgAcAF4AIABiwGIAcEEkgEDNi4xmAEAoAEBwAEB&sclient=gws-wiz
13. Medical News Today_Sites of UGIB. (n.d.). *Sites of UGIB*. Retrieved March 27, 2022, from https://www.medicalnewstoday.com/articles/upper-gi-bleed#:~:text=Upper%20gastrointestinal%20(GI)%20bleeding%20refers,and%20it%20can%20be%20serious.
14. Medline Plus_Hemorrhagic Stroke. (n.d.). *Medline Plus_Hemorrhagic Stroke*. Retrieved March 28, 2022, from https://medlineplus.gov/hemorrhagicstroke.html#:~:text=Hemorrhagic%20stroke%20is%20the%20less,artery%20wall%20that%20breaks%20open.
15. Miguel Coloma, P. (n.d.). *MINING ELECTRONIC HEALTHCARE RECORD DATABASES TO AUGMENT DRUG SAFETY SURVEILLANCE*.
16. Pacurariu, A., Plueschke, K., McGettigan, P., Morales, D. R., Slattery, J., Vogl, D., Goedecke, T., Kurz, X., & Cave, A. (2018). Electronic healthcare databases in Europe: Descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open*, *8*(9). https://doi.org/10.1136/bmjopen-2018-023090
17. Reference for EMA DARWIN-EU. (n.d.). *Reference for EMA DARWIN-EU*. Retrieved April 12, 2022, from https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu
18. Roberto, G., Leal, I., Sattar, N., Loomis, A. K., Avillach, P., Egger, P., van Wijngaarden, R., Ansell, D., Reisberg, S., Tammesoo, M. L., Alavere, H., Pasqua, A., Pedersen, L., Cunningham, J., Tramontan, L., Mayer, M. A., Herings, R., Coloma, P., Lapi, F., … Gini, R. (2016a). Identifying cases of type 2 diabetes in heterogeneous data sources: Strategy from the EMIF project. *PLoS ONE*, *11*(8). https://doi.org/10.1371/journal.pone.0160648
19. Roberto, G., Leal, I., Sattar, N., Loomis, A. K., Avillach, P., Egger, P., van Wijngaarden, R., Ansell, D., Reisberg, S., Tammesoo, M. L., Alavere, H., Pasqua, A., Pedersen, L., Cunningham, J., Tramontan, L., Mayer, M. A., Herings, R., Coloma, P., Lapi, F., … Gini, R. (2016b).

Identifying cases of type 2 diabetes in heterogeneous data sources: Strategy from the EMIF project. *PLoS ONE*, *11*(8). https://doi.org/10.1371/journal.pone.0160648

20. Schneeweiss, S., & Avorn, J. (2005). A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*, *58*(4), 323–337. https://doi.org/10.1016/j.jclinepi.2004.10.012
21. Souverein, P. C., van den Ham, H. A., Huerta, C., Merino, E. M., Montero, D., León-Muñoz, L. M., Schmiedl, S., Heeke, A., Rottenkolber, M., Andersen, M., Aakjaer, M., de Bruin, M. L., Klungel, O. H., & Gardarsdottir, H. (2021a). Comparing risk of major bleeding between users of different oral anticoagulants in patients with nonvalvular atrial fibrillation. *British Journal of Clinical Pharmacology*, *87*(3), 988–1000. https://doi.org/10.1111/bcp.14450
22. Souverein, P. C., van den Ham, H. A., Huerta, C., Merino, E. M., Montero, D., León-Muñoz, L. M., Schmiedl, S., Heeke, A., Rottenkolber, M., Andersen, M., Aakjaer, M., de Bruin, M. L., Klungel, O. H., & Gardarsdottir, H. (2021b). Comparing risk of major bleeding between users of different oral anticoagulants in patients with nonvalvular atrial fibrillation. *British Journal of Clinical Pharmacology*, *87*(3), 988–1000. https://doi.org/10.1111/bcp.14450