Concordance between primary and secondary electronic healthcare databases: A multi-database self-controlled case series study

Contents

1. Authors

Nicholas Hunt, Kanaka Soman, Patrick Souverein, Marloes Bazelier, Helga Gardarsdottir, Olaf Klungel

2. Summary

2.1. Lay Summary

Diagnoses are often recorded in different ways, depending on whether general practitioner or hospital data is used. It can be that data recorded from these two settings do not match in timing or whether there is a record altogether. When researchers would like to carry out an epidemiological study to assess, for example the side-effects of a medication, the recording of the diagnosis is used to determine the risk of the side-effect. Should this recording not exist in a certain data source or have an incorrect date, then bias may be introduced into the study, making the results less valid. In this study, we will describe the agreement in the recordings of major bleeding between the two healthcare data settings, in two European countries, the United Kingdom and the Netherlands. In particular, the existence of the recordings, the timing of the recordings and whether any recordings occur after the recorded death date. We will carry out a study assessing the association of major bleeding events identified from either healthcare settings and the use of direct oral anticoagulants or vitamin K antagonists, in a self-controlled study design. This will better inform decision making when designing epidemiological studies in electronic healthcare data.

2.2. Technical Summary

There is often mismatch between the recording of diagnoses in primary and secondary electronic healthcare data. Differences may exist in the recorded date of the event or whether it is recorded at all. For example, around two-fifths of all recorded stroke events are in both UK primary and secondary healthcare databases (within 120 days of each other) and around half of these had same-day recordings. The lack of concordance between different electronic health care records, which capture the same population, could lead to outcome misclassification and therefore bias, depending on which data domain is correct and then used in the epidemiologic study. Here we will describe the concordance between primary and secondary electronic healthcare data in the United Kingdom and the Netherlands in the occurrence of major bleeding. Agreement between the data settings, time gap between recordings and occurrence of recordings after recorded death date will be assessed. We will also compare the outcomes identified from different healthcare settings when applied to a self-controlled case series (SCCS) study. This will assess the association of major bleeding and use of direct oral anticoagulants or vitamin K antagonists for atrial fibrillation patients. The incidence rate of the outcome in exposed versus non-exposed time (incidence rate ratio) will be assessed, comparing outcomes derived from the different data domains. The aims of this study are to better inform pharmacoepidemiologic decision making.

3. Introduction

3.1. Rationale

The recording of events in primary and secondary electronic healthcare databases is often mismatched, whether that is the instance of the recording or the misclassification of the concept or timing of the event. Taking the UK as an example, around two-fifths of all recorded stroke events are in both UK primary (CPRD) and secondary healthcare (Hospital Episode Statistics) databases (within 120 days) and around half of these had same-day recordings.[1] Myocardial infarction was found to be recorded in both 51% of the time, while 8% of MIs were identified from disease registries but not in either primary or secondary healthcare data.[2] Fatal events are often best recorded in death registries and less so when using only electronic healthcare databases.[2] A lack of concordance between data sources could lead to misclassification and therefore bias, depending on which data source is most accurate. Events may be incorrectly misclassified in time periods of exposure and non-exposure, potentially impacting effect estimates.

3.2. Study Type

Methodological, pharmacoepidemiology

3.3. Objectives

Objective 1: Describe the concordance between primary and secondary care data in both the United Kingdom and the Netherlands

1. Determine the agreement in recording between diagnoses recorded in the primary and secondary care settings
2. Determine the time gap between diagnoses recorded in primary and secondary care, including time between a recorded sign, symptom and a confirmed diagnosis and potential death date

Objective 2: Compare the incidence of outcomes identified from primary and/or secondary care data in a self-controlled case series study (SCCS) design

1. Determine the risk of the outcome when using primary or secondary healthcare data
    a. H_0: The incidence of bleeding when using anticoagulants for atrial fibrillation will not be effect by the use of primary or secondary electronic healthcare data domains
2. Examine whether the assumptions of an SCCS are breached by using data from primary or secondary care data including:
    a. Event decreases probability of exposure
    b. Event increases probability of exposure
    c. No exposure can occur after event
    d. Event increases probability of death
    e. Non-independent event recurrences

This study aims to inform and therefore improve pharmacoepidemiologic methodology. This could have impact on future studies, thereby improving the assessment of drug safety and effectiveness potentially impacting clinical practice and policy.

3.4. Outcomes

- Percentage overlap of bleeding events occurring in the primary and secondary healthcare data domains within the same day, ±30 and 90 days.
- Percentage of bleeding events registered after death date.
- Overall incidence rates (IR) of major bleeding using primary and/or secondary care data.
- Incidence rate ratios (IRR) of major bleeding in the exposed time (first 30 days or including the remaining length of prescription) versus unexposed (baseline) time comparing primary and/or secondary care data.

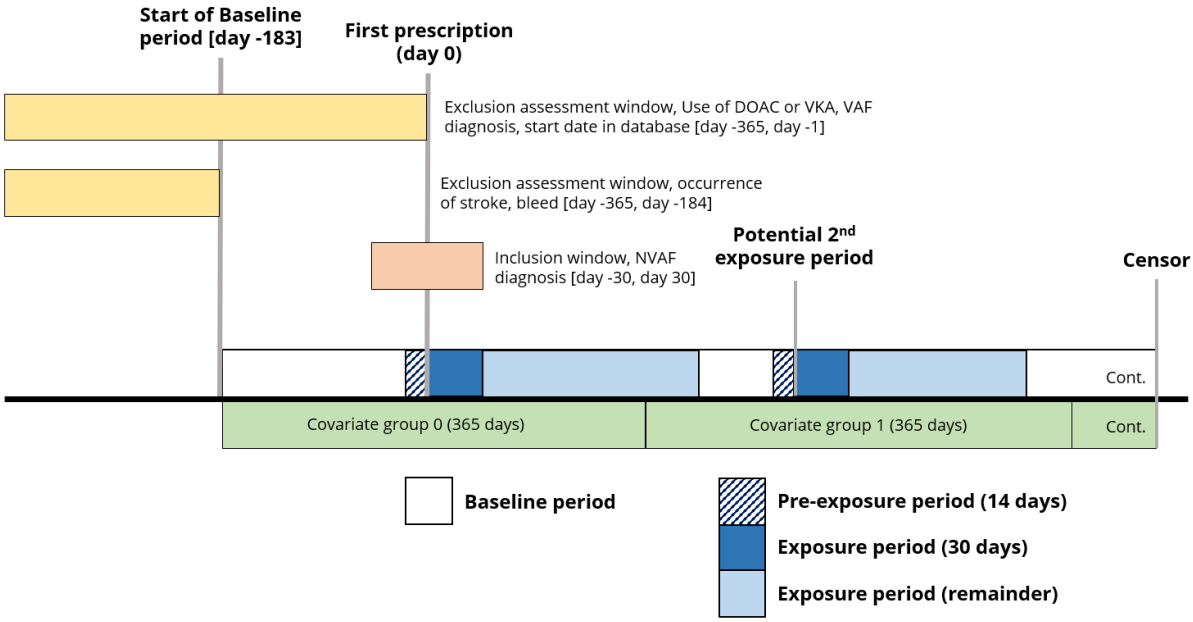
4. Methods

4.1. Study Design

For the first objective, we will determine the overlap of recording of a major bleeding event in the primary care data, after first identifying it from secondary care data. This ensures that only serious bleeding events are studied, that are eligible to be found in both. We will record the overlap of events that occur on the same day in the two data domains, as well as events recorded in the primary care data that occur within 30 and 90 days of the secondary care record. In addition, we will identify events in the two data domains that are recorded to occur after the recorded death date in the database.

For the second objective, we will use a self-controlled case series study design measuring the incidence of major bleeding when newly exposed to a DOAC or VKA compared to unexposed (baseline) period. In the study period 1st January 2010 to 31st December 2019 and using either primary or secondary electronic healthcare databases, as well as sourcing data from both. Patients will be censored at loss to follow-up, death, switching to the other drug class, or end of study whichever occurs earliest.

Figure 1. A graphical depiction of the study design.



To ensure the assumptions of the SCCS study design are not violated we will:[3]

- Study the first event only by starting the observation period 365 days into the follow-up time
- Carry out a sensitivity analysis to exclude cases where a death was reported within 90 days of the event
- Include a pre-exposure period which takes place from 14 days prior to the first prescription

4.2. Study Population

The study population will include those who initiated DOACs or VKAs therapy, aged ≥ 18 at the start of the baseline period, with a diagnosis of non-valvular atrial fibrillation (NVAF) ±30 days of initiation of anticoagulation therapy and occurrence of the outcome from either the primary or secondary EHR data in the study period 1st January 2010 to 31st December 2019. The index date is the date of first DOAC or VKA prescription. New users are defined as those who have not used either DOACs or VKAs in the 365 days prior to the first prescription.

Inclusion criteria:

- Presence in both primary and secondary EHR databases in the United Kingdom (CPRD Aurum) and the Netherlands (PHARMO Database Network)
- Aged ≥18 years at the first date of baseline
- Occurrence of major bleeding from either primary or secondary EHR databases in the study period
- Initiated DOAC or VKA therapy in the study period ±30 days of an NVAF diagnosis

Exclusion criteria

- Outcome had already occurred in the 182 days prior to the start of the baseline period
- History of valvular atrial fibrillation in the 365 days prior to the date of first prescription of a DOAC or VKA
- <365 days of observation time prior to the date of first prescription

4.3. Data Sources

CPRD Aurum (United Kingdom):

The Clinical Practice Research Datalink (CPRD) Aurum database consists of routinely collected electronic healthcare data from primary care practices in the United Kingdom, with the vast majority from England and Northern Ireland.[10] It captures diagnoses and symptoms, prescriptions prescribed by general practitioners, referrals and laboratory tests. In this study we will use Hospital Episode Statistics (HES) Admitted Patient Care (APC) data which contains all admissions to National Health Service providers in England. Although CPRD Aurum has accumulated over 41 million patients, around 35 million have a linkage to HES data.[4] CPRD Aurum diagnoses are coded to SNOMED CT (UK edition), while HES diagnoses are coded to International Classification of Diseases (ICD)-10.

The PHARMO Database Network (Netherlands):

The PHARmacoMOrbidity (PHARMO) Database Network consists of pharmacy dispensing data which can linked to other data sources including primary and secondary electronic healthcare data. Founded in 1999, it has accumulated to include 4.2 million active patients, which accounts for around 25% of the total Dutch population.[5] In this study, we used pharmacy outpatient data linked to primary and secondary care data. The primary care database accounts for around 20% of the total Dutch population and codes diagnoses in International Classification of Primary Care (ICPC).[6] Secondary care data diagnoses are coded in ICD-9 and ICD-10.

4.4. Feasibility Counts

Using numbers from the study which our case study is based upon, with data collected between 1 January 2008 and 31 December 2015 (Table 1), we predict the number of patients we will see in this study (Table 2). These numbers are adjusted to the study length of 1 January 2010 to 31 December 2019. The exposure definition is slightly different as we additionally include the more recently developed DOAC, Edoxaban. Note: These estimates apply only to the CPRD GOLD data source.

Table 1: The number of patients included in Souverein et al. 2020

	All NVAF Patients	Major Bleeding Outcome
DOACs	5852	205
VKAs	33277	1352

Table 2: The predicted number of patients that will be included in this study.

	All NVAF Patients	Major Bleeding Outcome
DOACs	8000	300
VKAs	46000	1900

4.5. Sample Size Considerations

Number of outcomes (Table 2) will be sufficient for the methodological study. We will stratify per DOAC or VKA and not by individual medicinal product. We will only include a small number of confounders so sample size will not be compromised in the analysis. This study does not test a clinical hypothesis, large sample size is not necessary. To identify a 1.5-fold risk and using an exposed period of 90 days, at an alpha of 0.05 and a 365 day study period, 232 subjects will be required with both a recording of a bleed and anticoagulant usage.

4.6. Planned use of linked data

CPRD Aurum will be linked to HES APC (secondary care) data. PHARMO outpatient pharmacy data will be linked to primary and secondary care data.

4.7. Selection of Controls

The study is self-controlled so exposed periods are compared to unexposed (baseline) periods.

4.8. Exposure

Using a SCCS design, periods of exposure will be compared to periods of non-exposure (baseline). Periods of exposure will be calculated using treatment episodes, while non-exposure is the period before or after the treatment episode. There are two separate exposures, DOACs and VKAs and risk of the outcome will be assessed separately due to the self-controlled study design. Any persons who switch between the two exposure groups will be censored from the study at the date of first new exposure.

Prescription Length:

We will assess exposure duration in a hierarchal manner: First we will consider the prescribed/dispensed number of tablets and the prescribed dosage. If this information is unavailable, then we will consider the median time between prescriptions per individual and based on ATC code, for the exposure to be the duration of use for a single prescription. This cannot be applied in situations where there are <3 prescriptions available, or if the estimated duration of the exposure >100 days (calculated by defined daily doses, DDD multiplied by the number of packages). In these situations, the mode of the estimated prescription duration for a particular drug in the total study population will be applied. Code lists for the included exposures can be found in a table in Appendix 1.

Treatment Episodes:

The follow-up time will be split into periods of either exposed or baseline time. Patients cannot switch between DOACs or VKAs, they will be censored at the date

of the prescribing/dispensing of the new medication. These treatment episodes will be constructed independent of possible dose changes within a period. The treatment episodes will be created to allow a 30-day gap between the theoretical end date of one prescription and the start date of the next prescription. Any overlapping days of the prescription, where the same drug is collected before the theoretical end of the previous prescription will be added to the end of subsequent prescription up to a maximum of 90 days.

4.9. Outcomes

The primary outcome will be the occurrence of any type of bleeding and is defined by the International Society on Thrombosis and Homeostatic as a symptomatic bleeding in an organ or other critical area. Major bleeding will include haemorrhagic stroke/intracranial bleeding (IC), traumatic intracranial bleeding, gastrointestinal bleeding (GI) and other unclassified extracranial bleeding events. These will be defined by SNOMED, ICD-10 (CPRD Aurum) and ICD-9, ICD-10 and ICPC (PHARMO) codes. These codes were identified and used in previous work and can be seen in Appendix 2.[7]

4.10. Covariates

Concomitant medications considered as potential confounders will be those that increase bleeding risk. The code lists can be found in Appendix 1:

- Non-steroidal anti-inflammatory drugs (NSAIDs)
- Corticosteroids
- Selective Serotonin Reuptake Inhibitors (SSRIs)
- Antiplatelet drugs

Comorbidities considered to be risk factors for bleeding. The code lists can be found in Appendix 2:

- Prior stroke/TIA, Pulmonary embolism (PE)
- Deep vein thrombosis (DVT)
- Hypertension
- Diabetes mellitus
- Cardiovascular disease (including congestive heart failure, angina, myocardial infarction, coronary artery disease, aortic plaque and peripheral artery disease)
- Alcoholism
- Liver disease
- Chronic kidney disease

4.11. Data/Statistical Analysis

The baseline characteristics will be stratified by treatment group (DOAC or VKA) and by data source (CPRD Aurum or PHARMO). The baseline period is defined as the unexposed reference period 30 days prior to use of a one of the exposures and unexposed time begins 30 days after the last calculated exposure. Means, standard

deviations (SD) and (percentage) totals will be calculated. Median follow-up will be calculated per treatment group in each data source.

Incidence rates (IRs) for events occurring within exposed and unexposed intervals will be calculated, along with incidence rate ratios (IRRs) comparing these two periods. The IRR and corresponding 95% confidence interval (CI) will be calculated using conditional Poisson regression. Time-varying confounders which are associated with the exposure and the outcome, such as age, will be accounted for in the adjusted model. The analysis will be stratified by sex (effect modifier).

Sensitivity analysis

In an SCCS, if an occurrence of the outcome leads to censor (e.g. death) then it breaches the an SCCS assumption. We will restrict the SCCS analysis to exclude persons who died in the study period.

We will assess the incidence of bleeding for the total exposed period and additionally only the first 30 days of each treatment episode. This allows the determination of risk period to be assessed and can inform whether there is potential exposure misclassification from construction of the treatment episodes.

Plan for Addressing Confounding

Potential time-invariant confounding (sex, genetics) will be accounted for by the use of the self-controlled study design itself. Several of these confounders are often unmeasured in electronic healthcare data. The study design removes much of the variation between individuals with disease risk.[8] Measurement of time-varying confounders (age, comorbidities, comedications), will account for instances of repeated or sustained exposure, such as with the use of anticoagulants. Comorbidities and comedications will be measured in time groups of 365 days, starting at the first day of the baseline period (day -183).

Missing data

Many confounders are addressed using the self-controlled study design and as such missing time-invariant confounder information will not impact the study. No missing information on exposure status is expected. Systematically missing outcome information is expected through the use of different data domains.

5. Data management

The CPRD Aurum data is stored within a secure environment at Utrecht University. The PHARMO data is secure within the PHARMO secure environment and access is provided remotely. Only (protocol) approved researchers will have access to the data. All documents will be archived within Utrecht University. Quality assurance of the data management will be ensured by the co-authors.

6. Patient or User Group Involvement

Patient groups will not be engaged in this study. The aim of this study is methodological and therefore patient involvement is not relevant. Identification of

the relevant study population exposure definition and outcomes are based on already published literature.[7]

7. Plans for disseminating and communicating study results

The study results will form part of a manuscript that will be published in a peer-reviewed international scientific journal. The results will also be presented at a national and/or international (pharmaco-)epidemiology conferences.

There are no restrictions on the extent and timing of the publication. The authors have no conflicts of interest to declare.

8. Ethical considerations

This is an observational study with no patient involvement so there are no requirements for ethical committee or institutional review board approval.

9. Limitations

The SCCS study design has some inherent limitations including potential breach of the assumptions leading to invalid and biased estimates including selection bias. When assessing exposure, there are assumptions that have to be made to estimate the treatment duration (episodes). The available data on exposure will be utilised in the methods described in 4.8 but there still may be a risk of information bias.

The observational nature of data can be a limitation while conducting this study as we use the routinely collected healthcare data whose primary aim is to provide care to patients and not research. Additionally, the use of multiple data sources from different countries is a limitation as there are differences in the way data are collected and recorded in these two databases due to the inherent differences in the healthcare systems of UK and the Netherlands.

10. References

1. Morgan, A., Sinnott, S. J., Smeeth, L., Minassian, C. & Quint, J. Concordance in the recording of stroke across UK primary and secondary care datasets: a population-based cohort study. *BJGP Open* **5**, 1–11 (2021).
2. Herrett, E. *et al.* Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* **346**, (2013).
3. Petersen, I., Douglas, I. & Whitaker, H. Self controlled case series methods: an alternative to standard epidemiological study designs. *BMJ* **354**, i4515 (2016).
4. CPRD Aurum HES APC January 2022 | CPRD. https://cprd.com/cprd-aurum-hes-apc-january-2022.
5. PHARMO database network. http://pharmo.nl/PHARMO-database-network/.
6. Kuiper, J. G., Bakker, M., Penning-Van Beest, F. J. A. & Herings, R. M. C. Existing Data Sources for Clinical Epidemiology: The PHARMO Database

Network. *Clin. Epidemiol.* **12**, 415 (2020).

7. Souverein, P. *et al.* Comparing risk of major bleeding between users of different oral anticoagulants in patients with on valvular atrial fibrillation. *Br. J. Clin. Pharmacol.* (2020).

8. Grosso, A. *et al.* Use of the self-controlled case series method in drug safety assessment. *Expert Opin. Drug Saf.* **10**, 337 (2011).