Development and validation of novel data science tools for quantifying drug exposure using routinely collected data

Alex Douglas Marshall

BSc (Hons), MSc

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

Institute of Health and Wellbeing

College of Medical, Veterinary and Life Sciences (MVLS)

University of Glasgow

June 2020

Abstract

Routinely collected healthcare data are increasingly being used as a source for research into the effectiveness and safety of drugs. Although these data have great potential, they require preparation before they can be used in research, a process which, amongst a number of other tasks, typically involves quantifying patients' exposure to the drug of interest. The aim of this thesis was to develop and validate a set of flexible, reusable functions for generating common drug exposure variables based on routinely collected prescribing data.

Six main classes of method for quantifying drug exposure were identified through a review of pharmacoepidemiological research; ever use vs. never use, use at a specified time point, daily dose or duration, persistence and discontinuation, adherence, and population level measures. The information obtained on these methods, their applications and the potential variations within each class formed the basis for developing an R package, prescribeR, which contains a range of functions designed to simplify and standardise the generation of drug exposure variables, and to provide a structure for reporting how these variables were produced.

The utility of the package was then demonstrated by applying it to two exemplar clinical studies, using a cohort of 5,571 patients with epilepsy constructed using linked data within the NHS Greater Glasgow and Clyde Safe Haven environment. In the first, prescribeR was used to quantify persistence to anti-epileptic drugs over the first 365 days of follow-up for cohort patients in order to assess differences in persistence across different drugs, as well as to compare persistence in new and existing users and patients prescribed monotherapy and combination therapy. All of the required persistence measurements for this study were generated using the prescribeR package, highlighting the relative ease of generating exposure data for a large cohort of patients and a number of different drugs.

In the second, the package was used to examine the effects of adjusting drug exposure definition on the estimated number of patients exposed to various drugs, the estimated exposure durations. The association between levetiracetam exposure and all-cause mortality was estimated using a range of time-fixed and

time-varying exposure definitions, and a wide range of hazard ratios and significance levels were observed across the resulting models, highlighting that the selected definition of drug exposure can potentially have a large impact on the results observed in clinical research.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to thank my supervisors Colin McCowan and Jill Pell for all of their support and guidance throughout my project, for looking over countless incomplete drafts, for reassurance and advice when I wasn't sure I knew what I was doing and for helping to make sure the last three and a half years have been a constantly interesting and enriching experience.

I'd also like to thank the following people and groups:

- The Medical Research Council for funding this project
- Craig Heath for giving me the opportunity to assist with his research, for allowing me to use his project data to test my work, and for his advice and support during the final year of my PhD
- My annual review panel, George Baillie and Olivia Wu, for providing invaluable feedback every year on my progress
- My examiners, Elizabeth Williamson and Jim Lewsey for conducting my viva and providing their feedback, and the convenor Daniel Mackay for arranging and chairing my viva
- The academic and teaching staff at the University of Glasgow, University of Edinburgh and the Karolinska Institute who lead the taught modules and training I have undertaken during my PhD, all of which helped me gain new skills which were essential in the completion of this project
- Kirstin Leslie, for verifying my systematic review, and also for being excellent moral support before, during and after every supervisor meeting over the last three and a half years
- The staff at the Robertson Centre, the Institute of Health and Wellbeing and the Safe Haven and Safe Haven support teams

- The administrative staff who have assisted throughout the course of my PhD, particularly Alexis Merry, Lorraine Waddell and Dorothy Brand

- The MVLS college librarian Paul Cannon for his assistance in structuring my systematic review and his advice during the early stages of my project

- Paul Mori from the university counselling service for his support, advice and guidance during a period where I was feeling very overwhelmed

- My friends, for always helping to keep me on track and motivated and for never failing to brighten my day. In particular I'd like to thank Kate for listening to me vent my frustrations every time something wasn't working, and for whisking me away for coffee and a movie every time I needed a break from one of those spirals

- My entire family for constantly being encouraging and supporting throughout all of my academic endeavours over the years. I'd like to especially thank my mum (Lynn), my dad (Douglas) and my brother (Mark) for putting up with me whenever I escaped back to Grangemouth to rest and recharge – they've often had to take the brunt of my frustrations throughout this process, especially during the last 3 months while we've all been cooped up during the lockdown, and I love them more than I can ever express

- Finally, I'd like to thank my dog, Judi, for always being happy to go for a walk when I needed to clear my head, even if it was the middle of the night

Author’s Declaration

I declare that the contents of this thesis are my own work and have not been submitted for any other degree at the University of Glasgow or any other institution. Where the work of others has been used it has been indicated and appropriately referenced.

Alex Marshall

Date: 15th June 2020

Definitions/Abbreviations

A&E	Accident and Emergency
AED	anti-epileptic drug
ATC	Anatomical Therapeutic Chemical classification
BNF	British National Formulary
CHI	Community Health Index
CI	confidence interval
CMA	continuous measure of medication acquisition
CMG	continuous measure of medication gaps
CNS	central nervous system
CPRD	Clinical Practice Research Datalink
CRAN	Comprehensive R Archive Network
DD	daily dose
DDD	defined daily dose
DID	DDD per 1,000 inhabitants per day
EDIS	Emergency Department Information System
EEG	electroencephalogram
EU	European Union
GP	general practitioner
GUI	graphical user interface
HR	hazard ratio
ICD-10	International Classification of Diseases, 10th Revision
IDE	Integrated Development Environment
IQR	inter-quartile range
ISD	Information Services Division
LOS	length of stay
MPR	medication possession ratio
NDC	National Drug Code
NHS	National Health Service
NHSGGC	National Health Service Greater Glasgow and Clyde
NICE	National Institute of Health and Clinical Excellence
NRS	National Records of Scotland
OMOP	Observational Medical Outcomes Partnership
PDC	proportion of days covered
PDD	prescribed daily dose
PIS	Prescribing Information System
PRE2DUP	prescriptions to drug use period
RCTs	randomised controlled trials
SIMD	Scottish Index of Multiple Deprivation
SMR	Scottish Morbidity Record
UK	United Kingdom
US/USA	United States/United States of America
WHO	World Health Organisation

1 Introduction

Drugs are one of the key forms of intervention available to physicians for the prevention and management of diseases.(1) Data from the 2014 European Health Survey found that 48.6% of participants across all of the EU member states reported having taken a prescribed medication in the two weeks prior to the survey – rates of medication consumption varied between the individual member states, ranging from 22.8% in Romania to 60.2% in Belgium.(2) The survey also reported that medication use was higher in older patients, with 21.9% of participants between 15 and 24 years old taking at least one prescribed medication compared to 87.1% of patients aged 75 years and over across all of the member nations. A study investigating the prevalence of use of medications amongst the elderly in the US reported similar results, showing that 88% of a group of 2,245 elderly patients with a mean age of 71 years were taking at least one prescribed medication.(3) In the United Kingdom (UK), prescribing is the most common patient-level intervention, and the second highest area of National Health Service (NHS) spending, below only staffing costs.(4) In 2016, the NHS in England dispensed 1.1 billion prescription items within a community setting, a 46.8% increase on the number of items dispensed in 2006.(5) These figures highlight the ever-increasing importance of drugs as a medical intervention, and therefore the importance of understanding the safety and effectiveness of these drugs.

1.1 Randomised controlled trials

Randomised controlled trials (RCTs) are considered to be the gold standard methodology for generating evidence on the efficacy and safety of medications.(6) The strength of RCTs as a source of evidence rests on the fact that they are carefully structured and controlled – a well-defined study population coupled with randomisation to treatment arms helps to minimise the effects of possible confounding and maximise internal validity, therefore theoretically providing an understanding of how effective the study medication is under ideal conditions. Large-scale RCTs are typically the final test of a drug's efficacy and safety in the drug development process, and the evidence generated is used to determine whether or not a drug is approved for release onto the market. RCTs are not without their limitations - they are typically extremely expensive to conduct, with one study reporting an average cost of between US$11.5 million and US$52.9 million to run a phase 3 clinical trial depending on therapeutic area.(7) The financial cost and organisational complexity of running large, multi-centre clinical trials limits the duration of follow-up, and therefore limits their ability to monitor long-term effectiveness and adverse effects of the medication being studied. The long timescales associated with planning, running and analysing RCTs also mean that they are not always able to keep pace with clinical advances, changes in public policy or urgent and emerging healthcare crises. Additionally, although the highly selective criteria used to select patients provides very strong internal validity of results, the use of restricted study populations coupled with the limited follow-up duration of trials can limit the ability to identify rare or late adverse events.(8)

Strong internal validity also typically comes at the cost of external validity and generalisability to the wider population. (8) A number of published studies have investigated the extent to which trial subjects reflect the wider population who would be prescribed the drug once it reaches the market. One study assessed the proportion of patients with a diagnosis of major depressive episode who met a standard set of exclusion criteria for anti-depressant medication efficacy trials.(9) The study found that out of 3,119 patients with a current diagnosis of major depressive episode, 75.8% were excluded by one or more of the eligibility criteria, with the presence of comorbid, non-depressive disorders and the duration of the depressive episode excluding the largest percentage of these individuals. A second study which assessed a cohort of 4,811 Scottish women with breast cancer using the eligibility criteria from twelve trials which have influenced national treatment guidelines found that while 73% of the women were eligible for at least one of the trials, the proportion of women from the cohort who were eligible for individual trials rarely exceeded 45%.(10) A review of 52 studies assessing the external validity of cardiology, mental health and oncology RCTs based on the representativeness of the study population showed that in 72% of the included studies the trial samples were not representative of the general population.(11) A number of the included studies found that RCT patient populations had a lower risk profile than real-world populations due to the exclusion of elderly patients and patients with comorbid conditions. These results help to highlight limitations on how well the results of safety and efficacy testing under trial conditions reflect the safety and effectiveness of these drugs in patients who would be receiving them in ‘real-world’ treatment scenarios after they have been approved.

A review of sources of evidence for decision making in healthcare discussed the fact that RCT results are often given additional weight when considering the available evidence, at the cost of other valuable data sources, and that the limitations discussed above mean that there is a need for other methodologies which can compensate for the limitations of RCTs, highlighting that no study design is flawless and that using complementary methods can help improve the quality of the overall evidence base.(12)

1.2 Pharmacoepidemiology and pharmacovigilance

Pharmacoepidemiology is the study of the use and effects of drugs in large, defined populations, adapting methodologies used for the study of disease in general epidemiology and applying them to aspects of clinical pharmacology.(13) Pharmacovigilance is the science related to the monitoring, understanding and prevention of adverse drug events with the aim of improving the understanding of medicines and patient safety.(14) The need for long-term studies of medications is highlighted by a number of cases where drugs have been approved for use and made available on the market, then have been subsequently recalled due to issues around safety that have only been discovered once they have been in general use. Perhaps the most well-known example of this is thalidomide, a drug which was prescribed to treat morning sickness during pregnancy in the late 1950s and early 1960s, and was subsequently found to cause a high rate of severe birth defects.(15) More recent examples include mibefradil, a drug used to treat hypertension and angina which the manufacturer withdrew a year after approval due to a range of possible drug interactions, some of which were potentially fatal,(16) and sibutramine, a weight loss drug which was removed from the market due to an increased risk of cardiovascular events and stroke in a subset of patients prescribed it.(17) A review examining patterns of post-marketing withdrawal worldwide identified 462 products withdrawn from the market post-approval attributed to adverse reactions between 1953 and 2013.(18)

1.3 Routinely collected data

As electronic systems have become more commonplace in the healthcare sector, an ever-increasingly large amount of data are being recorded about the care patients are receiving. As these data represent the process and outcome of healthcare provision in real-world practice, they have become a highly valuable resource for pharmacoepidemiology and pharmacovigilance. Routinely collected healthcare datasets can typically be classified as either administrative or clinical depending on the original purpose for which the data were collected.(19) Administrative data include databases containing information originally collected for the purposes of payment or reimbursement of healthcare services, as well as data recorded for service planning, and government registers of births, deaths, hospitalisations or disease.(19) Clinical databases, on the other hand, are typically records made during the provision of care by physicians or other members of medical staff containing data such as vital statistics, patient history, laboratory tests, diagnoses and procedures most often in the form of electronic hospital charts or general practitioner (GP) records.(20)

Both types of data can be useful to researchers, with the type of data being used determining the types of research questions that can be answered. Typically, routinely collected administrative data have good coverage, but collect limited, generic information whereas clinical systems may not have full coverage and can vary between geographic sites but will contain more detailed, clinically relevant information. So, for example, medical records taken from a general practice or hospital may provide a record of all drugs prescribed to patients by the doctors at that site as well as diagnostic and demographic information which could be useful for pharmacoepidemiology and pharmacovigilance but may not provide the full picture if a patient has also been treated at other sites whose records are not available to researchers. Administrative records taken from a pharmacy or insurance provider on the other hand would contain information on all drugs dispensed to patients from a range of sites, which could be useful for studies on prescribing patterns and drug utilisation but may not contain the detailed clinical information required to understand reasons for prescribing or outcomes of treatment.

The potential benefits of using routinely collected healthcare data for research include access to large, unselected populations with long-term follow-up, which can provide opportunities for greater power to investigate uncommon adverse events and to investigate the long-term effects of taking a medication.(21) As these data reflect how drugs are prescribed and consumed in ‘real-world’ settings as opposed to tightly controlled clinical trial conditions, there is also the potential to study interactions between different drugs being taken for comorbid conditions, and to study sub-groups of the population who are typically excluded from participating in clinical trials such as children, older people and pregnant women. The inclusion of these groups helps to improve the generalisability of the results of studies using routinely collected data to the whole population.

1.3.1 Routinely collected data in Scotland

The use of Scottish data for this type of research is of particular interest as Scotland has a comparatively non-mobile population, a single unified health provider responsible for delivering the majority of care (NHS Scotland) and a high incidence of a number of disease groups of interest including cardiovascular disease and a number of mental health conditions.(22) In Scotland, everyone who is registered with a GP is assigned a Community Health Index (CHI) number which is a unique identifier recorded during all of their encounters with NHS services.(23) The purpose of the CHI is to ensure that patients can be correctly identified at the point of care and that relevant information on the patients’ health is available to health services, but it also provides a method of linking routinely collected data covering hospital admissions, laboratory tests, births, deaths and prescribing with relative ease for use in research.(24)

All medications prescribed, dispensed and reimbursed within the community setting by NHS Scotland services are recorded in the Prescribing Information System (PIS).(25) This includes prescriptions written by GPs, nurses, dentists, pharmacists and a number of non-medical prescribers. It does not include prescriptions dispensed within hospitals but does include those prescribed by hospitals but dispensed at pharmacies in the community. PIS contains aggregate information on over 1.6 billion prescriptions reimbursed in the community nationwide from April 1993 onwards. Incorporation of the CHI number into PIS records, with complete coverage from 2009 onwards, means that individual level information is also available for over 507 million items prescribed and 344 million items dispensed from 2009 onwards.(26) Around 100 million new data items are added to PIS annually.(25) PIS has great potential as a data source for use in pharmacoepidemiology and pharmacovigilance as it contains data which are representative of the full national population and through linkage to other nationally collected healthcare datasets, can provide information on patient characteristics and long-term clinical outcomes.

1.3.2 Limitations of routinely collected data

Although there are advantages to using routinely collected healthcare data for research, these data are not without their limitations. It is important to consider the completeness and accuracy of these datasets before using them. Data completeness refers to how much of the required or expected data are present within the dataset; this can be measured by assessing the quantity of missing data.(27) Data accuracy describes how well the data correctly represent the 'real-world' events they describe.(28) A number of the limitations within these datasets are related to the fact that they were not initially created and collected for research purposes. Datasets may not contain all of the variables of interest for a particular research question, there may be differences in the way that routine data are coded and the level of detail available when compared to primary research datasets, or there may be errors in the data due to improper data entry.(29) Additionally, individual datasets may not contain all of the records relating to the patients who appear in the data. For example, if a country has a mix of private and public healthcare, or a number of different healthcare providers then the records of one provider may not contain a full record of each patients' medical history.(30) Some of these issues can be overcome by using data from a number of different sources, but this is not always straightforward. For example, different coding systems may be used across different sources, there may be difficulty in accurately linking subjects across datasets if common identifiers are not used, and there could be issues with combining the data on a technical level if the formats that data are stored in are different, particularly if proprietary software is used by the data controllers.

A number of studies have been conducted aiming to validate the information held in databases such as pharmacy records and compare them to other sources such as home medication inventories and patient questionnaires. One study which assessed the level of agreement between Medicare claims data, self-report and medication inventories for lipid-lowering drugs found that although there were instances where data did not match, there was generally a high level of agreement, with 86.5% overall agreement between medication inventory and Medicare claims data and 84.7% overall agreement between self-report and claims data.(31) A similar study which compared pharmacy data against self-report for patients taking medication for osteoporosis concluded that there was good agreement between the two sources, but that care had to be taken when attempting to define current use based on pharmacy data.(32) Overall, there is evidence to support the validity of routinely collected healthcare data as a source of information on which drugs patients have available, but these studies still stress that care should be taken in the assumptions that are made when using the data. It is also important to note that, even if there are complete records showing that a physician has prescribed a medication and a pharmacy has dispensed it, we cannot be sure whether or not a patient has actually taken the medication, or if they have taken it according to the dosage instructions.

1.4 Data preparation

Understanding and accounting for these potential limitations is essential when using routinely collected data for research, as identifying and accounting for potential errors in the data are essential in minimising the risk of bias. Bias is a systematic (or non-random) deviation of study results from the truth resulting from errors in measurements, selection processes, inferences, statistical analysis, or other procedures.(33) As described above, the fact that these data are not originally collected for research purposes can result in systematic issues with completeness and accuracy which can lead to bias if the data are used as collected.

Data preparation is an essential stage in any research project using these routinely collected data, as it provides an opportunity to address a variety of limitations of routinely collected data. It may not be possible to resolve all systematic issues within routinely collected data, but it is important that researchers address the known issues as much as possible during data preparation and analysis to maximise the validity of the results obtained. The exact process of preparing data for analysis varies, but typically involves exploration, cleaning and enrichment of the raw data. Before making any changes, it is important to explore the data to understand the characteristics of the dataset and identify potential sources of error. This will often involve understanding the processes by which the data are collected and the way that records are generated, the generation of summary tables or plots, and manual, unstructured exploration of the data available to help familiarise the user with its content and structure. Data cleaning typically involves steps such as removing irrelevant or duplicated data, dealing with missing, incorrect and implausible values, standardising data types, formatting and syntax, and verifying the data after making these changes.(34) Once the data are cleaned, steps can be taken to further enrich the data through processes such as the merging or linking of datasets from different sources or the generation of additional variables based on existing data.

Exposure to medication is usually the key exposure of interest in pharmacoepidemiological and pharmacovigilance research. Unlike in RCTs, however, detailed information on patient's drug exposure status is not typically recorded in routinely collected datasets. Therefore, drug exposure variables must be generated based on data such as records of dispensed prescriptions, hospital or GP records or insurance claims. There is no gold standard methodology for defining drug exposure based on routinely collected data. Different methods of quantifying drug exposure will provide varying levels of detail on individual patients' exposure status – for example, it is possible to split patients based on whether or not they have ever been prescribed a drug and compare them to those who have not or to assess changes in exposure status over the duration of a study period based on the quantity of a drug prescribed.

When deciding how to quantify drug exposure in any study, there are a number of factors which must be considered to minimise the risk of introducing either time-related or measurement bias and maximise the validity and accuracy of both the exposure variable itself and any other results obtained based on this measurement of exposure. To that end, it is important to take into account which exposure quantification methods are suitable for both the study design being used and the available data.

Studies using exposure data can largely be split into descriptive studies, where exposure itself is amongst the main factors of interest, and studies trying to estimate a causal effect, where the focus is on the effects of exposure on an outcome of interest. In studies where the focus is on a causal link between exposure and an outcome care has to be taken to separate the time period in which exposure is measured from the follow-up period where the outcome can occur in order to avoid immortal time bias. Immortal time bias occurs when the start of the follow-up period is incorrectly defined and a period of immortal time where the study event cannot occur is included during the follow-up period. This immortal time can lead to over-estimating the effectiveness of the treatment of interest. Exposure misclassification is a form of measurement bias which occurs when the measurement of exposure to the drug of interest used in the study does not accurately affect the reality of the patient's medication exposure. In order to minimise exposure misclassification in pharmacoepidemiology, researchers must select an exposure definition provides enough detail to answer the research question being investigated but can be accurately estimated based on the available data.

1.5 Aims

The primary aim of this thesis was to develop and validate a set of flexible, reusable functions for generating common drug exposure variables based on routinely collected prescribing data. This was split into three research objectives:

- Identify common methods used to quantify drug exposure based on routinely collected data through a systematic review of the literature

- Develop an R package containing functions for generating drug exposure variables based on the methods identified

- Test the package and demonstrate its utility in two exemplar clinical studies using real-world prescribing data on a cohort of epilepsy patients –

one aiming to measure rates of persistence to anti-epileptic medications and a second investigating the impact of varying drug exposure definition on observed outcome-exposure associations and a second

1.6 Thesis structure

This thesis consists of six chapters. This chapter has provided a general introduction to the use of routinely collected healthcare data for research, the strengths and limitations of these data and the need for data preparation. Chapter 2 describes a systematic review of pharmacoepidemiology and pharmacovigilance research aimed at identifying and classifying common methods for quantifying drug exposure using routinely collected data. Chapter 3 describes technical aspects of data preparation and exposure quantification, and provides documentation of the R package, prescribeR, which consists of a range of functions developed for generating drug exposure variables using the methods identified in the review. Chapter 4 discusses the data cleaning processes involved in the construction of a cohort of patients with epilepsy using linked data from NHS Greater Glasgow and Clyde and the use of the prescribeR package to examine rates of AED persistence amongst cohort patients during the first year of follow-up. Chapter 5 describes the use of this cohort to assess the impact that changing the definition of drug exposure has on the estimated number of patients exposed to a range of drugs and the rates of persistence to these drugs, as well as how using these different measures in statistical models impacts the association between levetiracetam exposure and all-cause mortality. Finally, Chapter 6 contains a summary of the findings of this thesis alongside a discussion of the limitations, potential for future work and final conclusions.

2 Review of methods to quantify drug exposure using routinely collected healthcare data

2.1 Aims

As previously described, drug exposure is typically the key exposure variable in pharmacoepidemiological and pharmacovigilance research and can be defined in a number of different ways based on routinely collected data. The main aim when generating drug exposure variables is to accurately describe which patients were exposed to the drug of interest and when in order to minimise the potential for bias in the form of exposure misclassification and to therefore maximise the validity of the study results. The aim of this systematic review was to identify and classify methods previously used to quantify drug exposure based on a variety of different routinely collected healthcare data sources.

2.2 Methods

2.2.1 Search strategy

The Ovid interface was used to search Medline and Embase to identify pharmacoepidemiological studies conducted using routinely collected health data. An initial basic search was performed using terms commonly found in key papers from the field. This was then expanded using synonyms and other terms found amongst the results of the basic search to give the complete list of search terms seen in Table 1, below. The search structure was reviewed by both a college librarian and PhD supervisors to ensure there were no key terms missing.

2.2.2 Review process – inclusion and exclusion criteria

Inclusion was limited to papers published in English between January 2012 and April 2017 (i.e., within the previous 5 years at the time of the initial search). Studies were considered eligible if their methodology included the use of at least one routinely collected healthcare data source to quantify subjects' exposure to one or more drug(s) of interest, or if they discussed the development or validation of methods of quantifying drug exposure based on these data. Studies were excluded if they exclusively used primary data collected directly from patients, however studies which used both routinely collected data and primary data for follow-up, comparison or validation were included. Conference abstracts and review articles were excluded.

#	Embase	Medline
1	pharmacoepidemiology/	Pharmacoepidemiology/
2	drug utilization/	Drug Utilization/
3	pharmacoepidemiology.tw.	pharmacoepidemiology.tw.
4	drug utili?ation.tw.	drug utili?ation.tw.
5	methodology.ab.	methodology.ab.
6	1 or 2 or 3 or 4 or 5	1 or 2 or 3 or 4 or 5
7	prescription/	Drug prescriptions/
8	electronic prescribing/	Electronic Prescribing/
9	pharmacy data*.ab.	Prescriptions/
10	prescri* data*.ab.	pharmacy data*.ab.
11	pharmacy records.ab.	prescri* data*.ab.
12	pharmacy claims.ab.	pharmacy records.ab.
13	electronic prescri*.ab.	electronic prescri*.ab.
14	administrative data*.ab.	pharmacy claims.ab.
15	7 or 8 or 9 or 10 or 11 or 12 or 13 or 14	administrative data*.ab.
16	exposure.ab.	7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15
17	drug exposure.ab.	Medication Adherence/
18	medication exposure.ab.	exposure.ab.
19	exposure period.ab.	drug exposure.ab.
20	adherence.ab.	medication exposure.ab.
21	compliance.ab.	exposure period.ab.
22	persistence.ab.	adherence.ab.
23	medication compliance/	compliance.ab.
24	drug exposure/	persistence.ab.
25	16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24	17 or 18 or 19 or 20 or 21 or 22 or 23 or 24
26	6 and 15 and 25	6 and 16 and 25
27	26 and 2012:2017.(sa_year).	26 and 2012:2017.(sa_year).

Table 1 - List of literature search terms and limits used to discover papers in Medline and Embase

2.2.3 Data extraction

Data were extracted from the eligible articles relating to country, type of database, use of data linkage, primary aim, study population size, population sub-group(s) investigated, drug/drug group investigated, and the method(s) used to classify drug exposure. Studies were assigned to four categories based on their primary aim (Table 2) adapted from a review of pharmacoepidemiological studies conducted using the Nordic prescription databases.(35) Studies which covered areas from multiple categories were assigned to the one which best described the primary aim of the paper. Drugs of interest were mapped to the relevant British National Formulary (BNF) chapter.(36) Databases were classified as either administrative or clinical and studies were considered to have used linked data if they connected records at individual patient level across multiple data sources.

	Description
Drug effect	Studies into the efficacy of drug therapies, including studies aiming to validate or extend the results of randomized controlled trials in either whole or sub-populations and studies investigating the efficacy of approved drugs for off-label or alternative uses
Drug safety	Studies related to the negative outcomes associated with exposure to a drug, including the incidence of death, hospitalisations, adverse reactions and drug interactions
Drug utilisation	Studies investigating issues around prescribing, dispensing and consumption of medications including trends in prescribing, effects of changes in public policy, health economics and patient compliance with prescribing instructions
Validation /methodology	Studies whose main aim was to discuss methodological aspects of using routinely collected data to quantify drug exposure, or validating these databases as a source of information

Table 2 - Descriptions of the primary aim categories used for studies included in this review (35)

2.3 Results

A summary of the number of papers found, excluded and retained at each stage of the review process can be found in Figure 1. A total of 1,267 papers were identified from Medline and Embase, 31 of which were duplicates. The review of titles and abstracts identified 1,008 ineligible results – these included conference abstracts, studies which collected primary data prospectively using case report forms or questionnaires, data which were repurposed from clinical trials instead of routinely collected data, and papers which focused on medical events other than prescribing such as exposure to bacteria.

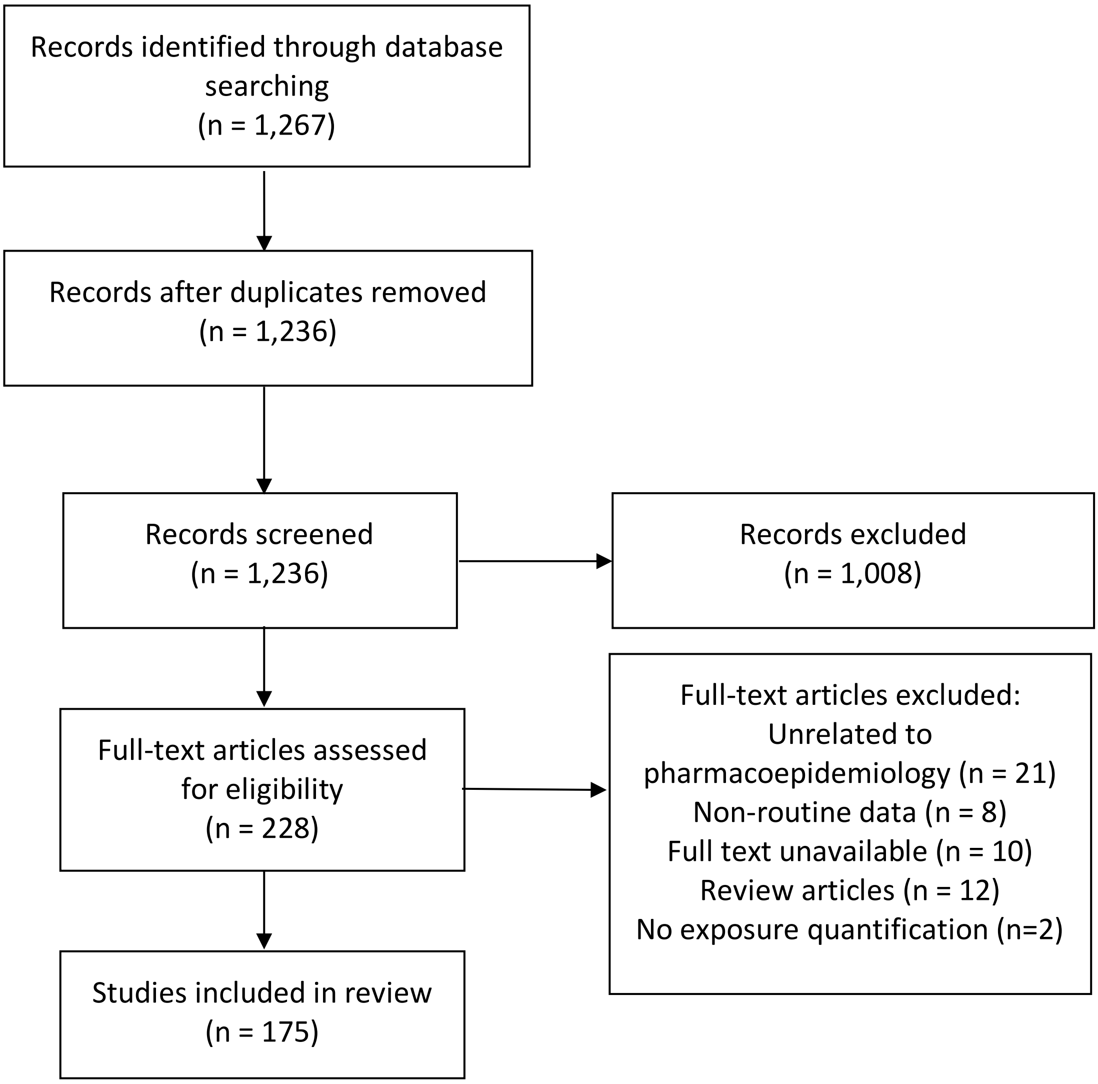

Records identified through database searching
(n = 1,267)
Records after duplicates removed
(n = 1,236)
Records screened
(n = 1,236)
Records excluded
(n = 1,008)
Full-text articles assessed for eligibility
(n = 228)
Full-text articles excluded:
Unrelated to pharmacoepidemiology (n = 21)
Non-routine data (n = 8)
Full text unavailable (n = 10)
Review articles (n = 12)
No exposure quantification (n=2)
Studies included in review
(n = 175)

Figure 1 – PRISMA flow diagram (37) detailing the studies included and excluded from this review

The full text of the remaining studies was retrieved for review, and the relevant information extracted, as described above, after which papers were deemed to eligible or ineligible for inclusion. A further 53 studies were removed during this stage, resulting in 175 articles eligible for inclusion in the review.

2.3.1 Exposure assessment

The methods section of each included study was reviewed to determine how exposure to the drug(s) of interest was quantified, and the methods were grouped into the classes shown in Table 3. Some papers used multiple methods to quantify exposure, so were included in more than one group.

	Description	Publications N (%)
Ever use vs. never use	Study classified patients as exposed or unexposed based on whether or not they had at least a threshold number of prescriptions for the drug(s) of interest)	108 (61.7)
Use at time point	Study classified patients as exposed if they were prescribed the drug of interest at a specified point in time, or within a given time frame based on an index or outcome date	112 (64.0)
Daily dose/ Duration	Study defined periods of exposure based on prescription dates and some form of daily dose method or prescription duration	97 (55.4)
Persistence/ Discontinuation	Study defined length of treatment persistence by defining points of discontinuation based on the length of time between prescriptions for the drug of interest	61 (34.9)
Adherence measures	Study estimated subjects' adherence to medication using measures such as medication possession ratio or proportion of days covered	45 (25.7)
Population level	Study assessed exposure at a population level, considering changes in measures such as number of prescriptions, DIDs or prevalence of medication use over time	53 (30.3)

Table 3 - Summary of methods used to quantify exposure to drugs in the included studies, n = 175

The simplest method of classifying drug exposure was dichotomising patients based on whether or not they have ever had at least a threshold number of prescriptions, or ever versus never use. This method was commonly used to define inclusion and exclusion from a study population, or to define groups within a population. An example of how patients are split using this method can be seen in Figure 2 - the first two patients have at least 1 prescription for the drug of interest during the follow-up period and are classified as exposed, whereas the third is classified as unexposed. Using this method, no distinction is made between the first two patients even though they have different total numbers of prescriptions.

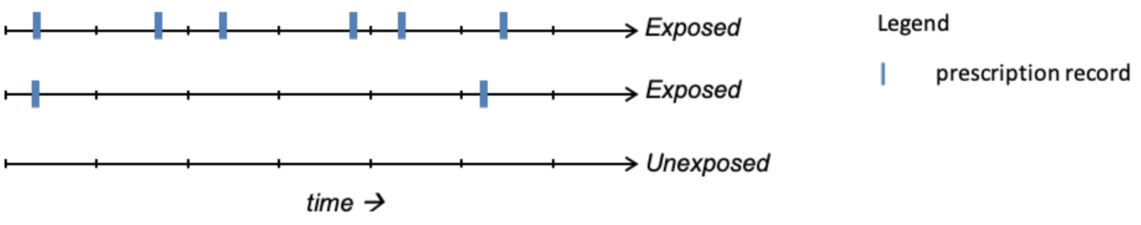


Figure 2 - Example application of the ever use methodology

The second class of methods involved classifying patients based on whether or not they had been prescribed a drug of interest within a specified time frame. Common examples of this method in the reviewed studies included classifying patients as new or existing users of the drug of interest at baseline based on whether or not they had prescriptions for the drug of interest before their index date (Figure 3), and classifying patients as current or past users of the drug of interest at an event such as death or hospitalisation based on whether or not they were prescribed the drug of interest within a set number of days before the event (Figure 4).

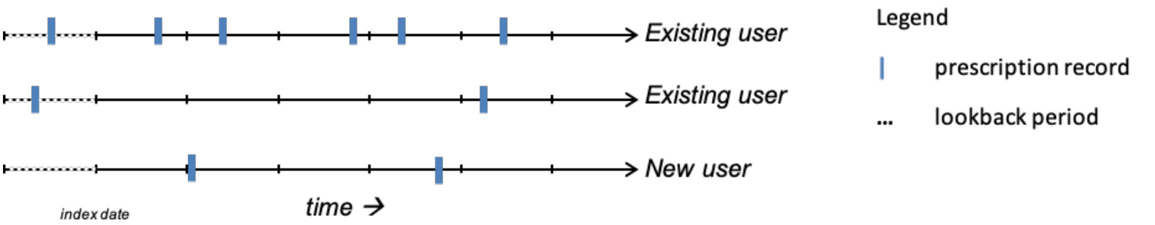


Figure 3 - Example application of a use at time point definition to split patients into existing or new users at baseline

The examples shown in Figure 3 and Figure 4 highlight the technical similarities of these methods; in both cases exposure is defined based on the presence or absence of prescriptions for the drug of interest adjacent to the event of interest.

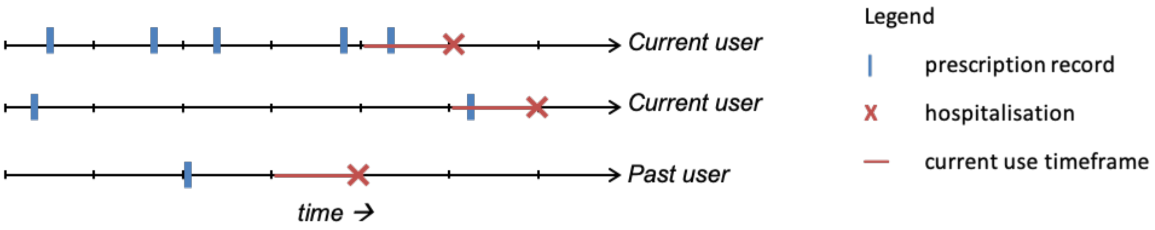


Figure 4 - Example application of a use at time point definition to split patients into current or past users of a drug at the time of a hospitalisation

Another example of a method assessing use at a specific time involved splitting follow-up into time windows and checking for prescriptions within these windows. Example applications of this method included defining exposure as a continuous variable by splitting the whole follow-up period into windows and assessing exposure during each window separately and defining maternal exposure to medications during the three trimesters of pregnancy when assessing the risk of later birth defects in their children. In the example shown in Figure 5, the patient would be considered exposed during the first and third windows but not during the second.

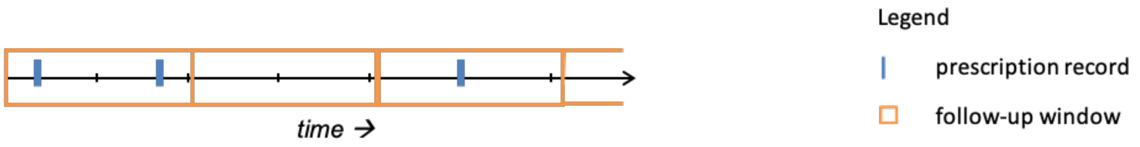


Figure 5 - Example of defining exposure to a drug of interest within consecutive windows during follow-up

Both ever use and use at time point methods were used in a number of studies to define exposure to additional drugs which were not the main focus of the study as a categorical variable for use as a covariate in statistical modelling, a study characteristic, or a proxy for an outcome or comorbid condition. For example, one study examining the prevalence of multimorbidity in a population of Scottish patients used prescriptions for relevant medications as a method of identifying patients with pain conditions, depression, anxiety, migraine and psoriasis.(38)

Daily dose (DD) methods went a step further, defining periods of drug exposure by assigning expected durations to individual prescriptions based on the amount of a drug prescribed. The example patient timeline shown in Figure 6 demonstrates how a duration of exposure is applied to individual prescriptions to define periods where the patient theoretically had access to the drug of interest. The specific method used depended on the type of data available. Some studies used a days' supply variable included in their data or calculated a prescribed daily dose (PDD) based on the amount prescribed and dosage instructions. In the absence of individual instructions, other measures such as the defined daily dose (DDD), the number of tablets dispensed or a fixed duration for individual prescriptions were used. The DDD is the dose of a drug defined by the World Health Organisation (WHO) as 'the assumed average maintenance dose per day for a drug used in its main indication in adults'.(39) Some studies also used the total number of DDDs prescribed to a patient as a measure of cumulative exposure to a drug during follow-up.

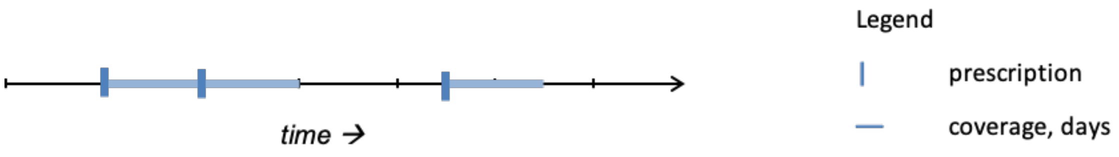


Figure 6 - Example of defining duration of exposure using a daily dose method

Another method used to define the duration of exposure was calculating persistence. Persistence can be defined as the length of time between when a patient is first being prescribed a medication and when they stop (or discontinue) treatment.(40) Persistence was typically measured using the refill gap method, where persistent time is defined by assessing the gaps between successive prescriptions, defining discontinuation as any point where the time between prescriptions exceeds an allowable gap. Figure 7 demonstrates how prescription records are connected using this method to define periods of persistent use and points of discontinuation.

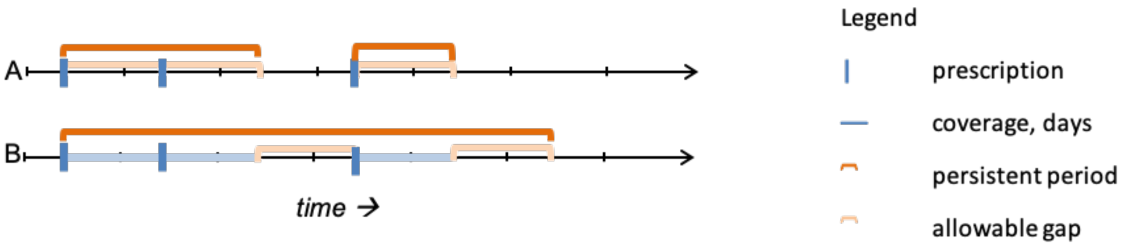


Figure 7 - Example application of the refill gap method for defining persistence without (A) and with (B) the coverage of individual prescriptions taken into account

Some studies combined persistence and DD methods, applying the allowable gap from the assumed end of prescribed supply instead of the prescription date. Some studies also used the date of the first prescription of another drug of interest, or medication switching, as an additional discontinuation point. The example shown in Figure 7 highlights how taking prescription durations into account alters the periods of persistent use observed when using the same patient data and allowable gap length.

Adherence methods describe the extent to which a patients' medication-taking reflects instruction.(41) The most commonly used adherence calculations were the medication possession ratio (MPR) and the proportion of days covered (PDC), both of which calculate the percentage of time where the patient had medication available. Figure 8 below illustrates the concept of calculating adherence using these methods. MPR is calculated as the ratio between the total days' supply and the length of the observation period. The MPR can overestimate adherence for patients who regularly refill their medications early, resulting in a value greater than 1. The PDC is similar conceptually, but more conservative. It is calculated by defining the number of days which were 'covered' by medication and dividing by the total number of days. Oversupply from one prescription is moved to any gaps between supplies instead of accumulating, and the PDC is inherently capped at 100%. In most of the included studies, subjects were considered to have adhered to their medication if their adherence value was greater than 80%. A less commonly used method which is mathematically similar to MPR is continuous medication acquisition (CMA), calculated as the total supply divided by the length of the observation period. Some studies measured adherence using the continuous measure of medication gaps (CMG), which determines the number of days where there was no supply and divides this by the length of the observation period.(42) By definition, calculating adherence using any of these methods requires the use of one of the DD methods described above to define the expected supply of the medication of interest.

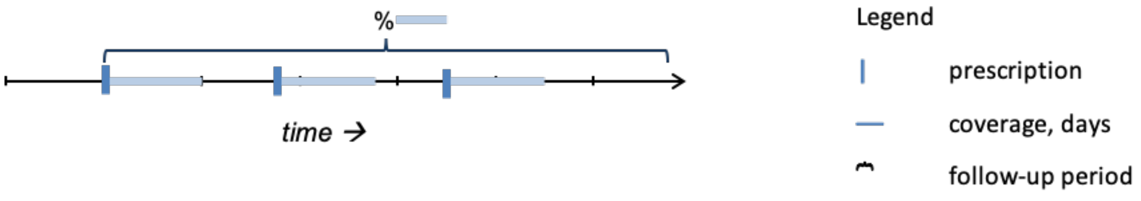


Figure 8 - Example patient timeline demonstrating how adherence is calculated using individual prescription records

The final class of methods focused on population-level exposure to a drug rather than individual patients' exposure. These studies were often interested in the changes in prescribing over time, and reported measures such as the number of prescriptions, the number of DDDs per 1,000 inhabitants per day (DIDs), the prevalence or incidence of prescribing of a drug of interest or exposed person-time across the population. An example showing how individual prescribing records can be used to determine population-level trends is shown in Figure 9

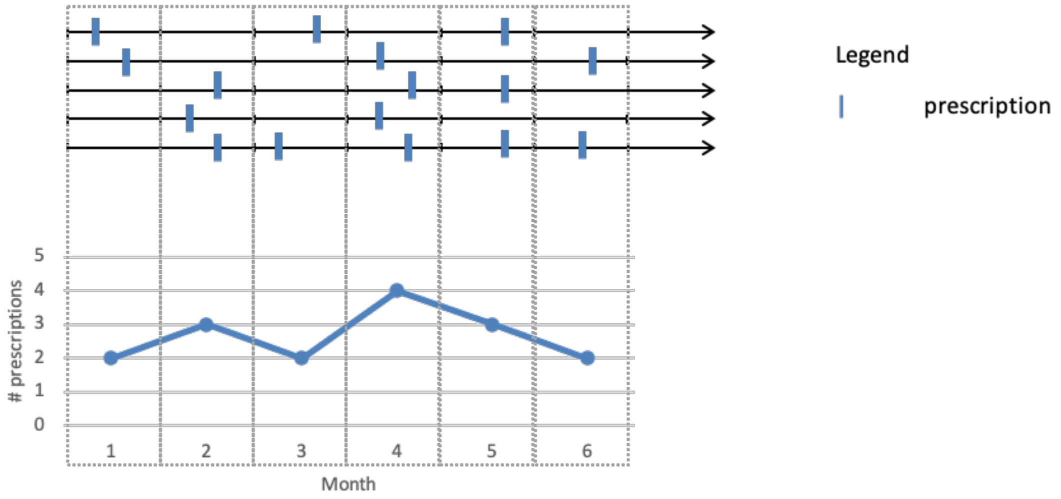


Figure 9 - Example of how individual prescription records can be used to measure population level exposure

Examples of common variations on the methods within each category and how these were applied with references to representative studies can be found in Table 4.

Exposure category	Method	Definition	Ex.
Ever use	Ever use vs. never use	≥ n prescriptions of drug X	(43)
	Number of drugs prescribed	no. drugs/classes with ≥ n prescriptions reported	(44)
	Number of prescriptions	no. prescriptions of drug X reported	(45)
Use at time	Use within period of interest	≥ n prescriptions of drug X within n days of specified date	(46)
	New user selection	0 prescriptions for drug X in the n days before index date	(47)
	Use within windows	≥ n prescriptions of drug X within consecutive periods of m days	(48)
	Concurrent use	≥ n prescriptions of drug X and drug Y within n days of each other	(49)
	Current/past use	drug X within n days before index date = current user drug X > n days before index date = past user	(50)
DD/ duration	Defined daily doses (DDDs)	Start date + DDDs dispensed	(51)
	Prescribed daily doses (PDDs)	Start date + PDDs dispensed	(52)
	Days' supply	Start date + days' supply dispensed	(53)
	Quantity prescribed	Start date + quantity dispensed	(54)
	Fixed duration	Start date + fixed n days per prescription	(55)
	Cumulative supply	Total no. DDDs dispensed	(56)
Persistence	Refill gap	Discontinued if no prescription after > n days from previous prescription	(57)
	Refill gap with coverage	Discontinued if no prescription after > n days from end of previous prescription duration	(58)
	First prescription to last	Persistent between the first and last prescriptions for drug X	(59)
	Switching	Drug X discontinued at date of first prescription for drug Y	(60)
Adherence	Medication possession ratio (MPR)	Ratio of total days' supply obtained to length of study period	(61)
	Proportion of days covered (PDC)	Number of days covered by supply divided by the length of study period	(62)
	Continuous measure of medication acquisition (CMA)	Total days' supply divided by the length of study period	(63)
	Continuous measure of medication gaps (CMG)	Length of study period minus the number of days' supply, divided by length of the study period	(64)

Exposure category	Method	Definition	Ex.
Population	Number of prescriptions	No. prescriptions of drug X per unit time	(65)
	DDD/1000 inhabitants/day (DID)	No. DDDs of drug X per 1000 inhabitants per day	(66)
	Exposed person-time	Total person-time exposed to drug X for population	(67)
	Prevalence/incidence	Proportion of new or ongoing users of drug X per unit time	(68)

Table 4 - Specific examples of applications, with references to studies which used those methods

2.3.2 Country

The key characteristics of the studies included in this review are described in Table 5. The most common country of origin for the data used in the included studies was the United States (US), which accounted for 29.1% of the included studies. Six of the included studies used data from multiple countries. Only countries which accounted for more than 5 of the included studies are included in Table 5 – there were also studies reviewed which used data from Australia, Finland, Germany, Korea, Japan, Serbia, Spain, Brazil, Croatia, Ireland, Israel, New Zealand, Oman and Switzerland.

2.3.3 Study aims

The most common primary focus of the included studies was drug utilization, which encompassed a broad range of different study topics. These included studies on topics such as rates of adherence, persistence, discontinuation or medication switching and the factors affecting these behaviours. Other drug utilisation studies focused on changes in prescribing following changes in health system policy or legislation, the impact of healthcare costs, varying insurance coverage and other socioeconomic factors on prescribing, and potential for abuse of prescription medications.

The validation and methodology studies were similarly varied in focus, including papers on topics such as methods for quantifying and avoiding exposure misclassification, the effects of restrictive insurance coverage, over the counter medications and medicine samples on data completeness, validation of databases compared to other sources of medication data, possible sources of confounding and methods for reducing its effect on results. There were also studies describing methods for enhancing routinely collected prescribing datasets, including methods for estimating prescription durations where this information is not present in the data, generating structured dosing information from free text records and ways to combine different databases to address methodological issues occurring when only one source is used.

2.3.4 Database type

Administrative databases were the most common type of source database among the studies included in this review – 82.9% of studies used at least one administrative data source and 73.7% of studies used them exclusively. Administrative data were either in the form of transactional records from private or national health insurance claims and data taken from prescribing or disease registries. The data available varied according to the origin country's healthcare system. The USA has a private healthcare system, so records here typically originated from the records of individual insurance companies or from records of government programs such as Medicare and Medicaid. In countries such as France, Japan, Canada and Taiwan, which all have national health insurance schemes data were typically obtained from reimbursement records for these schemes. Studies based in countries such as Denmark, Norway, Sweden and Scotland used data taken from prescribing registries.

Clinical data were less common than administrative data in the included studies - 26.3% of studies used clinical databases, and 17.1% used them exclusively. These data were mostly primary care records from networks such as the Clinical Practice Research Datalink (CPRD) or the Italian Health Search network, although some secondary care data were also used, including data from the Veteran's Administration in the USA. Linkage of multiple data sources was quite common in the identified studies, with 35.39% of studies using some form of record linkage. The most common example of this in the included studies was linkage between national prescribing registers and disease, birth and death registers, but there were also a number of studies which linked clinical and administrative data.

	Publications N (%)
Data Origin	
United States	51 (29.1)
United Kingdom	19 (10.9)
Denmark	18 (10.3)
Canada	12 (6.9)
Italy	10 (5.7)
France	8 (4.6)
Netherlands	8 (4.6)
Sweden	6 (3.4)
Taiwan	6 (3.4)
Multiple	6 (3.4)
Norway	5 (2.9)
Other (<5 studies per country)	26 (14.6)
Primary aim	
Drug effect (43, 49, 69-74)	8 (4.6)
Drug safety (46, 48, 50, 55, 56, 58, 75-112)	44 (25.1)
Drug utilization (44, 45, 47, 51-54, 57, 59, 60, 62, 63, 65-67, 113-192)	95 (54.3)
Validation/methods (61, 64, 68, 193-217)	28 16.0)
Database type	
Administrative	129 (73.7)
Clinical	30 (17.1)
Both	16 (9.1)
Linked data	63 (36.0)
Drug(s) of interest	
Gastro-Intestinal Drugs	4 (2.3)
Cardiovascular Drugs	31 (17.7)
Respiratory System	6 (3.4)
Nervous System	52 (29.7)
Infection	12 (6.9)
Endocrine System	31 (17.7)
Genito-Urinary System	1 (0.6)
Malignant Disease	4 (2.3)
Musculoskeletal System	8 (4.6)
Eye	2 (1.1)
Skin	1 (0.6)
Multiple	23 (13.1)
Study population	
< 999	20 (11.4)
1000 – 9999	37 (21.1)
10 000 – 49,999	49 (28.0)
50 000 – 99,999	10 (5.7)
100 000 – 499,999	23 (13.1)
> 500,000	21 (12.0)
Unknown/variable	15 (8.6)
Sub-group	
Children (aged <18)	7 (4.0)
Elderly (aged >60)	18 (10.3)
Pregnant women	13 (7.4)
N/A	137 (78.3)

Table 5 - Summary of characteristics of included studies, n = 175

2.3.5 Drug group

There was a wide range of drug classes investigated. Drugs acting on the central nervous system (CNS), including anti-depressants, anti-psychotics and other psychotropic drugs as well as anti-epileptics and opioids, were the most commonly studied, accounting for over 25% of all studies. Drugs affecting the cardiovascular and endocrine systems were also quite common, accounting for approximately 17% of studies each. Over 10% of the total number of studies investigated more than one type of medication – these were typically papers that investigated a range of drugs as part of a methodological study, papers that looked at patients' entire medication history as part of a drug utilization or safety study or papers that investigated overall rates of prescribing across populations of interest.

2.3.6 Study population size

There was a wide variation in study population size across the included studies, ranging from 80 participants to 29.5 million participants with a median of 22,152. Of the 175 studies included in this review, 55 had a population of more than 50,000 patients. The largest study had a population of over 29 million patients – this was an outlier, which made use of the entire dataset of a large US health insurance provider and studied utilization patterns of all drugs prescribed to these patients. Study population was not reported for 15 studies, either because these were conceptual methodology papers which did not apply the methods to a specific group of patients, or because they were population level studies where the study population varied over the study period.

2.3.7 Sub-groups

There were three main population sub-groups investigated across the reviewed studies – children, elderly patients and pregnant women. A number of different age cut-offs were used to define both children and elderly patients across these studies, so for the purposes of this review any paper that focused on patients aged 18 and under or 60 and over respectively were included in the relevant category. The studies of all three sub-groups focused on a range of different drugs, but around half of the studies in each sub-group were focused on CNS-active drugs.

2.4 Discussion

The aim of this systematic review was to identify the methods used to quantify patients' exposure to drugs in studies using routinely collected prescribing data. A range of common classes of methods for quantifying exposure were identified based on the methodologies of the included studies: ever/never use, use at a specified time point, DD-related, refill gap, adherence measures and population level measures.

2.4.1 Exposure quantification methods

These methods provided different types of information on patients' exposure to medication, and each has different advantages and potential sources of random error or bias. Ever use of a drug is easy to establish, even with limited data, but it does not differentiate between patients who are one-time or infrequent users of a drug and those who are long-term users. Some of the included studies calculated the total number of prescriptions for the drug of interest, but this is a simple measure of cumulative exposure and does not account for differences in the quantity prescribed or dosing. In some studies, thresholds other than one prescription were used. This may be clinically relevant and could help to select patients who are ongoing users of a drug, but still does not account for changing exposure status with time or cumulative exposure as shown in Figure 10, where all three patients meet the threshold for exposure but have different patterns of prescribing across the follow-up period. Despite these limitations, ever use can still be a useful method for defining a cohort, as it can establish a baseline of exposure which can be expanded upon through other methods.

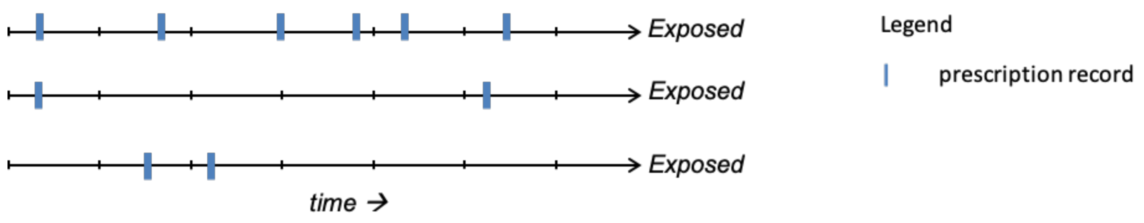


Figure 10 – Example of potential variation in patients who meet a 2-prescription threshold for exposure

Use at a specified time point can provide a more precise definition of exposure by focusing on exposure only during the time adjacent to the outcome of interest or by splitting users into categories based on recentness of use at time points of interest. This may provide more valid results, but these methods alone still do not account for variation in cumulative exposure. It is important to ensure that the timeframe of interest is set correctly to minimise exposure misclassification. When defining new use, shorter windows may misclassify existing users as new, which will affect the validity of the results obtained by blending different groups of users. Allowing for a longer window will increase the ability to identify genuinely new users of the drug of interest but could be overly selective and exclude patients who would be suitable for analysis or limit the size of the study population through the need to have data available for this lookback period. Similarly, when defining current use at the time of an event the timeframe of interest should be set to account for the amount of time where you would expect the drug of interest to affect the risk of the outcome. As with ever use, these methods can still be useful for defining cohorts, population sub-groups or additional study characteristics such as comorbidity, but care has to be taken that the thresholds set are specific enough so as to minimise risk of bias.

Creating more detailed exposure periods using a daily dose method provides even more detail, but the accuracy of these exposure periods will vary depending on which method is used to define the duration of each prescription. Using the number of DDDs dispensed can introduce bias, as it is based on the most common dose when used as a monotherapy for a specific indication. It may be suited to some patients, or drugs where there are a limited number of indications or dosage options, but it may not be an accurate reflection of many patients' treatment. Figure 11 shows the different lengths of exposure calculated based on two prescriptions for simvastatin when using the DDD of 30mg per day compared to a patient specific PDD of 10mg per day. In this example, using the DDD under-estimates the coverage of each individual prescription, leading to a higher proportion of unexposed time compared to the PDD.

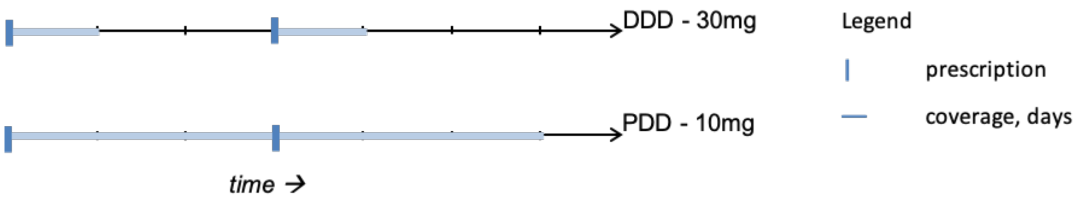


Figure 11 – Comparison of the duration of exposure to simvastatin calculated using the DDD and PDD

This is especially true for drugs with multiple indications, a wide range of potential dosages or where there is variation in dosing over time. Figure 12 shows an example of the potential difference in the length of exposure from 1 prescription for the anti-epileptic drug carbamazepine when using the DDD of 1g per day compared to a PDD of 200mg per day. In this example, the DDD overestimates the duration of the prescription significantly – if error of this nature is introduced for a number of similar prescriptions over the course of the follow-up period this could have a large impact on the results of a study.

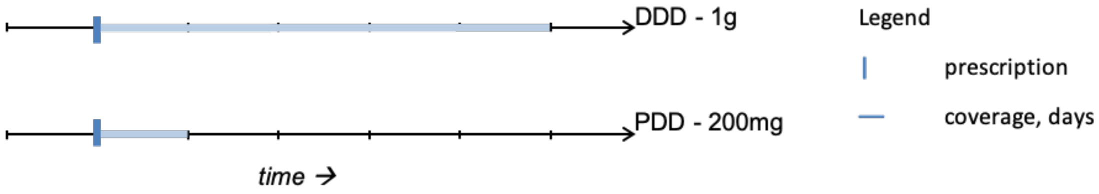


Figure 12 – Comparison of the duration of exposure to carbamazepine calculated using the DDD and PDD

One study which investigated concordance between the DDD and a pharmacist-defined value for days' supply for eight drug classes found that using DDDs to define exposure duration could lead to exposure misclassification of different magnitudes depending on drug class.(209) This is less of an issue when using days' supply or PDD as they are patient specific, but not all datasets include this information, and there is still room for error as we do not know whether patients follow instruction or take their medication.

Persistence and discontinuation also allow for variation in exposure over time. Here, the main potential source of bias is the definition of the allowable gap. Defining the allowable gap requires understanding of the usual clinical practice with respect to the country, drug of interest and the healthcare system the data represent. A study conducted by the Observational Medical Outcomes Partnership (OMOP) compared a 0-day and 30-day allowable gap and showed that a 30-day gap provided more clinically meaningful information.(218) Shorter windows may be more suitable in cases where medications are frequently reviewed by physicians, such as in cases where a narrow therapeutic window necessitates therapeutic drug monitoring. Where possible, combining daily dose and persistence may allow for more precision, as the allowable gap between prescriptions only needs to allow for a grace period between prescriptions for factors such as non-adherence, medication stockpiling and late refill of prescriptions. If no dosage information is available, larger allowable gaps are required to account for the supply of medication provided by each prescription. This will likely vary between patients or individual prescriptions, so the selected allowable gap may have to be larger to account for this variation in addition to the other factors listed above. Setting too long an allowable gap risks misclassifying genuinely non-persistent or unexposed time as exposed, which also introduces error. Figure 13 shows the potential variation in the observed persistent use periods based on two prescriptions depending on the methods used. In this example, both a 60-day allowable gap and a 30-day allowable gap between the end of the supply of one prescription and the next prescription are sufficient to define these as the same use period, whereas a 30-day allowable gap between prescriptions alone is not enough and results in two separate use periods.

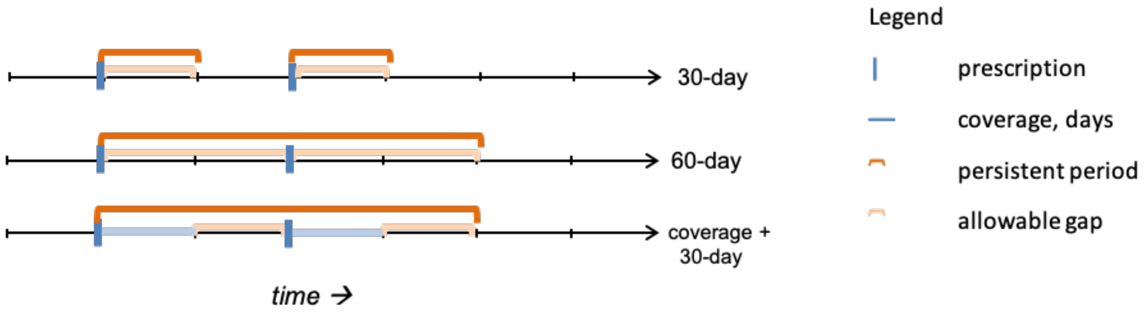


Figure 13 –Comparison of persistence measures using 30- and 60- day allowable gaps without coverage and a 30-day allowable gap with coverage

The potential for error in adherence measures is similar, in that errors can occur in both the MPR and PDC if the follow-up period or coverage from drugs supplied are incorrectly defined. There is also potential for error in studies that combine multiple observation periods, as early dispensing and high adherence in one period can offset late dispensing and poor adherence in another; giving a combined measure which does not accurately capture changes in the patient's medication availability.(42) Methods such as the MPR and CMA which use the sum total of medication dispensed over the study period can overestimate adherence if patients regularly refill early, and can return adherence values of greater than 1, or 100%. As described above, adherence measures typically depend on the use of a DD method to define the amount of a drug supplied to a patient, introducing further potential for error depending on the method used.

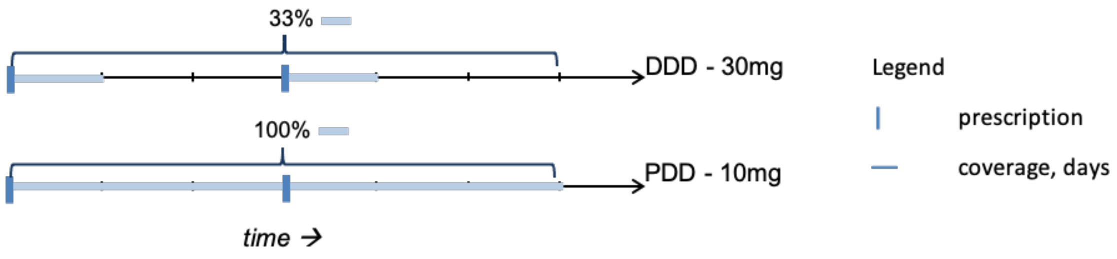


Figure 14 - Comparison of adherence values calculated based on the DDD and PDD from two prescriptions

In the example shown in Figure 14 using the DDD as opposed to the PDD would result in an under-estimation of drug supply, and in turn of the patient's level of adherence during the period of interest. Adherence and persistence methods were often used together, as measuring persistence can establish observation periods for adherence calculations. In these cases, errors made when defining the allowable gap will also affect the accuracy of the adherence measurement.

Although the type of information and level of detail clearly varies across the different methods, few studies assessed the impact of using different methods to classify exposure on the definition of the study population or explored the potential effect that using different methods could have on the observed associations between drug exposure and outcomes. One study which did assess this compared the effect of different assumptions of daily dosage and methods of accounting for gaps and overlaps between individual prescriptions on the construction of treatment episodes, and the effect these methods had on the observed association between SSRI exposure and hospitalisation. The results showed that assumptions of 1 DDD per day or 1 unit per day had a significant impact on the duration of exposed time compared to patient-specific dosages, and that this in turn gave rise to both over and under-estimations of risk associated with exposure to the drugs investigated depending on the method.(219)

A similar study investigated the effect of using time-fixed definitions of exposure based on different threshold numbers of prescriptions and time-varying methods based on different prescription durations on the association between use of oral anti-hyperglycaemic drugs and mortality. The results of the study demonstrated large differences in the observed association based on the classification of exposure used.(220) This study also highlighted the potential for immortal time bias in these time-fixed definitions in cases where the follow-up time was not defined in the same way for the exposed and unexposed populations. In this case, starting follow-up for both groups from an index prescription for any anti-hyperglycaemic drug introduced immortal time for the exposed patients between their index prescription and their first metformin prescription, which in turn

Taken together, these studies show potential for variation in study results based on the exposure definition chosen and highlight the need to consider which methods are appropriate for the dataset being used and for the study design. However, more evidence is needed to further clarify the potential for error across a wider range of drugs of interest and using different methods. It is also important to highlight that routinely collected data provide information on drugs that have been prescribed or dispensed to patients, they do not provide information on whether dispensed drugs were actually taken; therefore, none of the methods discussed avoid this source of potential exposure misclassification.

Although most of the reviewed studies typically contained enough information to categorise and summarise the methods used to generate exposure variables, many of the studies contained only limited detail regarding the structure of the database(s) being used or the computational processes involved in preparing the data for analysis. Since there is a great deal of heterogeneity across routinely collected databases, greater transparency when describing data exploration, cleaning and enrichment would give other researchers a clearer idea of the strengths and limitations of the published results, as well as enable easier reproduction of the study on other datasets. Increasing the reproducibility of research using routinely collected data is important, as validating the results of studies in different populations or using different databases would serve to further increase the impact of the results obtained. Additionally, transparency in the methods used may lead to increased standardisation in how certain aspects of data preparation are handled, which would reduce the complexity of the process and comparability of results across different studies. The RECORD statement outlines the key elements which should be reported in epidemiological research using routinely collected data, and advises that the algorithms and codes used to derive all of the exposure variables used should be provided where possible, so adhering to this or other similar checklists when publishing research can help ensure that the relevant information is included.(221)

2.4.2 Strengths

Although there are published studies discussing individual methods for defining exposure this literature review is one of the first studies to collect and summarise information regarding the range of methods used across the published literature in the field. A number of reviews have been published examining aspects of pharmacoepidemiological research using routinely collected data. For example, reviews have been published which summarise the research output from specific databases or groups of databases, including the Swedish Prescribed Drug Register,(222) the Nordic prescription databases,(35) and the German LRx database.(223) There have also been reviews published focusing on evidence related to specific outcomes and drugs of interest, including a review of studies investigating the possible link between use of thiazolidinediones and risk of bladder cancer(224) and a review of treatment adherence in patients being treated for headaches.(225)

As shown in Table 5, 16% of the studies reviewed had methodology or validation of routinely collected healthcare data as a primary aim, and although there was a general lack of evidence comparing the full range of different approaches to defining exposure, there are studies and reviews covering aspects of one or more of the methodological approaches discussed in this review. One study focused on exposure misclassification, using a probabilistic bias approach to identify misclassified categorical exposures. The proportion of days covered was calculated using a DD measurement, and this was used to assess the effect of misclassification on hypothetical exposure-outcome associations. The authors concluded that misclassification was an important factor to consider in pharmacovigilance, and that measures such as the one demonstrated were useful sensitivity tests for estimating potential bias.(193)

Other studies that were included in this review focused on other ways to expand and improve existing methods of assessing drug exposure. For example, one study proposed an extension to the PDC method of assessing adherence which took into account changes in adherence over time as opposed to assessing adherence as a single, time-fixed variable. The authors highlighted the potential for bias in the current method, and the benefit of a more accurate method for determining adherence in studies which associate adherence with a clinical outcome as well as those where having an understanding of how and why adherence changes over time is important.(200) A second study focused on the idea of modelling exposure as a time-dependent variable as opposed to a time-fixed one. The Prescriptions to Drug Use Periods (PRE2DUP) method aims to create a model of individual patient's exposure based on their prescribing history and the amount of the drug dispensed in DDDs among other factors. In a study aiming to validate the method by comparing the resulting model against patient interviews, the authors also highlight that an advantage of this method is that the core process of generating the exposure periods is iterative and takes into account the patient's common refill patterns when imputing missing data for individual prescriptions. This allows for the consideration of behaviours such as stockpiling and early refilling which were highlighted as potential reasons for exposure misclassification in other literature.(203, 226)

These studies discussing specific methods of quantifying exposure are part of a wider section of the literature relating to methodological issues in studies making use of routinely collected data. Understanding potential limitations specific to the data being used is an important part of the planning stage of any study, and a number of the reviewed studies cover methods for handling such issues. For example, there are a number of issues discussed across the included papers that should be considered in the case of health insurance claims data.

These can include unmeasured confounding factors, such as restrictive reimbursement policies, the use of free drug samples or out-of-pocket drug purchases. In each of these cases, it is possible that the patients' available supply of the medication of interest is under-estimated based on claims data alone, and if this is not accounted for there is potential for exposure misclassification, even if exposure is correctly quantified for the data that is present. One study assessing the impact of restrictive reimbursement policies in two Canadian provinces found that the administrative data available only captured 61% of the dispensations for the drug of interest where such a policy was in place, and concluded that in cases where no attempt was made to account for this the resulting bias could be quite large.(227) Another study which investigated completeness of US commercial claims records for warfarin found that there were cases where patients paying out-of-pocket for the generic form of the drug resulted in the potential for incomplete data in the claims database.(198) These issues are all specific examples relating to the use of claims data, and may not apply to other database types. Dispensing records from a pharmacy would likely be able to account for out-of-pocket prescription drug purchases but may be missing medications available from other sources. This is highlighted by a study of the rates of free prescription drugs based on data taken directly from office-based physicians. The study found that rates of free sample provision varied across drug type, and between branded and generic drugs, but that there was potential for bias in studies relying solely on pharmacy claims data.(228)

Even with an understanding of the database being used, the process of data preparation is usually one of the most time consuming parts of studies using this type of data.(229) Once data cleaning is complete, the process of data enrichment typically involves enhancing data either by deriving additional key variables using existing data or by linking multiple data sources. Two of the studies included in this review described the creation of structured data fields based on free text information encoded in prescribing records. In both cases, algorithms were created to recognise common phrases and numerical values in free text fields and convert this into standardised fields which could then be more easily used in further research using these databases.(230, 231) Development of a common data model was a large part of the work conducted by OMOP, with the main aim being to facilitate the use of data from multiple sources and enable reproducible drug safety research. This model took a person-centric approach, where drug exposure timelines were created in a standardised manner – two case studies were produced which demonstrated the potential to transform disparate source databases into comparable databases for use in research, with the researchers proposing that this meant that studies produced based on these data were more meaningfully comparable and therefore more able to appropriately impact public policy.(218) Another published study discussing the potential benefits of combining different sources for drug safety research highlighted that increasing the overall population size and diversity of the studied population would help increase validity of results. This study went on to discuss the need for data to share a common formatting to ensure ease of comparison of different records, but discussed that this can present a range of challenges, particularly where records from different countries are involved.(232) All of these potential issues should be taken into account when planning a study using routinely collected data, and these choices will impact the decisions made with respect to how data are cleaned and manipulated.

Another strength of this review is that the studies included were not limited to a single database or data from a single country. This is beneficial when considering methods used to determine exposure based on routine data, as the nature and quality of the data available differ from country to country, particularly where there are differences in the way healthcare is provided. These differences have an effect on the way the data has to be processed. An example of this can be seen in the difference between healthcare provision in the USA and in the UK, and the studies included in this review based on data from these countries. The USA has a private healthcare system, and most of the data used in research in US-based studies in this review was taken from insurance companies. As has been discussed above, data from insurance companies can have issues around completeness resulting from factors such as out-of-pocket drug purchases and restrictive coverage. Additionally, the data will likely represent a specific subset of the population who are able to afford insurance, or even patients who are from a specific industry if the insurance provider from whom the data were sourced provides insurance through patients' employers. Other sources of data commonly seen in US-based studies were data taken from specific benefit programs such as Medicare and the Veteran's Health Administration. As with insurance data, these sources will represent specific sub-populations as opposed to whole populations and will have data missing where drugs are not covered by the benefit program in question. In contrast, the UK has a mostly public healthcare system, with the NHS being the major healthcare provider. The most common database seen in the included studies published in the UK is the CPRD, a primary care dataset made up of medical records from a network of general practices in the UK which contains data on approximately 6.9% of the UK population.(233) As well as this primary care data, CPRD also offers linkage to other datasets such as secondary care and mortality records for around half of these patients. In contrast to the US datasets, there is not as much of a concern of fracturing the population based on coverage level as the NHS is the main healthcare provider, and CPRD data are broadly representative of the population as a whole.(233) However, most of the data regularly available is on primary care only, so there is still potential for missing data with relation to outcomes and prescribing in secondary care, as well as for medications purchased over the counter. The overall population size is also much smaller than the US, and so uncommon conditions, outcomes or drugs may be more difficult to study due to drawing subjects from a smaller population. In addition to these issues, there are likely differences in prescribing behaviours, legislation, approval status and public policy surrounding specific drugs. This comparison highlights a number of potential reasons that different methods would be useful depending on the source of data for a study, and in turn highlights the benefit of examining studies published internationally rather than limiting to one database or country in this review.

2.4.3 Limitations

This review is not without its limitations. Although a wide range of studies were found, this is not a comprehensive review of all research published using routine data, so it would therefore be invalid to draw conclusions about time trends in the amount of research published based on routine data, or the overall prevalence of certain study characteristics in the field as a whole. However, the main aim of this literature review was to characterise different methods being used in pharmacoepidemiological research to quantify or classify exposure, and a range of methods were identified in the papers that were found. Although the included studies were diverse, it is possible that some eligible studies were not identified from the search - expanding the search strategy to include other international database types such as ‘drug register’ may identify further eligible results. Additionally, the methods used to classify study characteristics in this review may not have adequately captured certain details of the included studies. Studies were only categorised by primary aim, when a number of studies could realistically belong to more than one category – for example, a number of drug safety studies also considered effectiveness of the drug of interest, and secondary components were relatively common in studies that were classified as having validation or methodology as a primary focus.

This review focused specifically on the use of routine healthcare databases to measure drug exposure. Therefore, the findings cannot be generalised to using similar data sources to address other research questions. Additionally, information was gathered on the methods used as they were reported, so there may be information missing if not enough detail was given by the authors as to how data were handled, or if steps involved in cohort generation were not described in the text. Finally, while the review identified the types of methods used and their frequency, a systematic comparison of their relative capacity to minimise exposure misclassification was outside the scope of the review.

2.4.4 Conclusion

Research conducted using routinely collected healthcare data has the potential to be a valuable source of evidence for real-world clinical practice, and there is potential for this type of research to complement some of the limitations of other methodologies such as RCTs to provide a more complete picture of how effective a wide range of drug treatments are. This review highlighted a range of different methods for determining patients’ exposure to medicines based on routinely collected prescribing data; however, further work is required to better understand comparative strengths and limitations of these methods in minimising exposure misclassification, and the impact that choice of exposure measurement has on the associations between exposure and outcomes observed in studies using these data. Additionally, further work is required to increase the level of transparency and reproducibility in the reporting of how routinely collected prescribing data are prepared for research in order to maximise the potential impact of results and evidence generated based on these data.

The next chapter will describe the development of a set of flexible, reusable R functions for generating drug exposure variables using the various methods identified in this review.

3 Documentation of R functions for quantifying drug exposure

3.1 Introduction

In the previous chapter, a review of published pharmacoepidemiology research identified a variety of potential research questions which can be answered using routinely collected data. These were split into four main categories; studies investigating the effectiveness of drug therapies, studies into the safety of medications, drug utilisation studies, and studies related to the methodology of using routinely collected data. Each of these study types required patients' exposure to the drug(s) of interest to be quantified during the data preparation process. A range of different methods for quantifying drug exposure were also identified and grouped into classes. These classes were ever use vs. never use, use at a specified time point, daily dose or duration, persistence and discontinuation, adherence, and population level measures.

3.1.1 Data preparation

As described in Chapter 2, data preparation is essential in studies using routinely collected data, as it provides an opportunity to address the limitations of the raw data and minimise the potential for bias in the subsequent analyses. Data preparation is typically one of the most time-consuming stages of studies using routinely collected data.(229) Manual exploration of the data are required to understand the content and structure of the data and identify specific issues which must be resolved. Data cleaning involves making adjustments and then verifying the validity of the data. Further inspection of the data is then necessary to verify these new values and identify any additional issues. Once the existing data are considered to be valid, they can be used to generate additional variables of interest which must also be verified. This entire process is iterative and can be particularly complicated when using especially large datasets, datasets which include variables stored as free text or when combining multiple datasets, as all of these factors will introduce greater variety to the data. Sources of error may not be initially obvious and may only become clear after some initial modifications such as standardising object types or correcting syntax issues. The need for manual data exploration and trial and error during data cleaning and enrichment mean that researchers often create code for data preparation on a project-by-project basis. This lack of standardisation between studies, combined with the variation in the level of transparency in the reporting of the specific processes used to transform raw data into analysis datasets can impact the reproducibility of research using routinely collected data.(234) While it is not feasible at present to create algorithms that can completely automate the process due to the potential complexity and variation of routinely collected data, it is possible to develop tools which will make it easier to standardise and repeat specific data cleaning or enrichment tasks.

3.1.2 Data preparation in R

An important consideration when working with routinely collected data is the choice of data management and analysis software. There are a wide range of options available, including programming languages such as SQL, R and Python and statistical software such as SAS, SPSS and Stata. When choosing software to use for research it is important to consider a number of factors including ease of use, functionality, flexibility and extensibility, cost and licensing issues, available sources of support or learning resources, scalability, integration with other software and the ability to produce output in the desired formats. For this project, the decision was made to develop a package for use in R.

R is a programming language and computing environment for data manipulation and analysis, calculation and graphics. It was developed by Ross Ihaka and Robert Gentleman as an implementation of the S programming language.(235) One of the key advantages of R when it was first released was that while S was only available as a commercial software package, S-PLUS, R was free, open source software. S, and by extension R, has its roots in data analysis, and is useful for both interactive exploration of data and programming to create new tools. R is typically accessed through the use of a command line interface. Commands to R take the form of expressions or assignments.(236) When an expression is given as a command, it is evaluated, and the result is printed and lost. Assignments are evaluated and then the results are passed to a variable and stored. Users can interact with R through the use of individual expressions, but one of the core benefits of R is that it is easy to build up to developing functions, essentially a series of commands executed in sequence. R is an object-oriented language, meaning that everything that is manipulated in R is an object, or a named data structure.(237) The simplest object type in R is a vector, which can contain a set of elements of the same basic data type (i.e. logical, integer, double, character, complex or raw). R has a range of basic object types, including matrices, arrays, lists and data frames, which vary by their dimensionality (1d, 2d or Nd), and whether they allow heterogenous or homogenous data types. Data frames are the most common format for storing data in R for data analysis. Technically, data frames are held in R as lists of equal-length vectors, meaning that they appear as matrices but are capable of storing data of different base types.

There are a number of reasons why R was chosen as the analysis software for this project. As mentioned above, R is open source and free to download, use and update, making it a cost-effective option for data analysis in research. The fact that it is free to use and available across a number of different common platforms and operating systems also means there is a lower barrier to entry when collaborating with others or when sharing code.(238) Since it is widely used in both academia and industry, there are a wide range of support and learning resources available, including books, training courses and tutorials and community forums where users can raise questions and share solutions to problems. Although it is possible to interface with R using its base GUI, the use of an integrated development environment (IDE) such as RStudio makes it easier to maximise the potential of R as an analysis tool and development platform. RStudio is designed to improve the utility of R, providing a more intuitive and user-friendly interface and tools such as syntax highlighting, code validation and completion, debugging, version control, environment management and package development tools.(239)

3.1.3 Packages in R

One of R’s major advantages is that it can be easily extended through the creation of new functions and packages. In R, a package is a collection of code, data, documentation and tests developed to expand the base functionality of R or add new capabilities. R is a powerful and flexible language, able to handle large amounts of data and complex calculations, and there are a wide range of packages available which build on R’s base data querying and manipulation capabilities, allow R to read data from a wide variety of formats, interact with other data analysis or statistical software and producing high quality, publication-ready outputs. There are currently over 15,000 packages available from the Comprehensive R Archive Network (CRAN) providing functions for a variety of data analysis tasks, including packages for use in healthcare research and data analysis.(240) This wide repository of packages provides a strong foundation for further expansion, as existing improvements to R’s data handling capabilities make it easier to develop tools for specific data analysis and manipulation tasks such as handling prescribing data.

Creating a package is a good way to share code with others, as it is easy for other R users to download, install and use packages. Packages can be a useful way of writing and storing code even if it just for personal use, as the standard conventions for writing and storing a package provide an effective structure for a project and make it easier to reuse or update the code at a later date.

Building a package containing functions for performing common data preparation tasks such as generating drug exposure variables has a number of benefits. As discussed above, R’s availability means that these tools can be easily distributed, can be used in conjunction with other R packages, and can be easily expanded if desired. When collaborating with others or publishing research, the methods used to generate exposure variables can be easily reported by referring to the packages and functions used, making it easier to understand how data have been processed or replicate the analysis. As mentioned above, improving the transparency and reproducibility of analyses makes it easier to validate findings using other datasets, which can, in turn, help to ensure that research findings are reliable and have the appropriate impact.

3.1.4 Aims

The aim in this chapter was to use the information gathered on the common methods for defining drug exposure during the systematic review to build an R package, prescribeR, which contains a variety of functions for quantifying drug exposure based on routinely collected prescribing records. The remainder of this chapter will go into detail on the contents of the prescribeR package. Although six classes of methods were defined in the previous chapter, the decision was made to initially focus on four when creating the prescribeR package - ever use, use at a specified time point, daily dose related methods and persistence. Adherence was excluded as there is already a published R package, AdhereR, which contains a set of functions for measuring and visualising adherence to drugs using routinely collected data.(241) Some of the population-level measures described in the systematic review are included in the prescribeR package's summary functions, but the decision was made to focus initially on the individual-level exposure methods as these were more commonly used in the studies included in the systematic review in Chapter 2.

The functions in the prescribeR package can be split into three major groups; data standardisation, data summaries and exposure classification, with the latter split further into ever use, use at a specified time point, daily dose and persistence. The remainder of this chapter provides the rationale for each group and sub-group as well as any necessary definitions, then describes the individual functions, including the R code, as well as a plain language breakdown of the data manipulation process and output, the arguments, and usage examples demonstrating different possible analyses using synthetic data.

3.2 Generation of synthetic datasets

The data governance in place to safeguard personal healthcare information aim to balance the need to maintain patient privacy and confidentiality while allowing for the controlled use and sharing of information for purposes such as improving care or medical research. Data provided to researchers are typically anonymised, and there are restrictions on the sharing of research data with third parties or the use of data for other purposes in order to minimise the potential of accidental disclosure of sensitive patient information. There is, however, a benefit to being able to show how the R functions described in this chapter interact with and transform data, so the decision was made to produce a synthetic dataset. The aim in constructing this synthetic dataset was to produce data which has the key characteristics of real prescribing data but does not contain any identifiable data from real patients and can therefore be shared or published alongside the exposure algorithms.

Even when data preparation is complete, datasets can contain records for large numbers of patients and a wide range of different drugs. Including a small synthetic dataset with a limited number of patients and drugs in this R package allows new users to easily explore the data, and then test different functions and analyses quickly and easily to understand the functionality of the package before using it on real data. It also provides a sample dataset which can be used for any further development and testing. With a small dataset it is easier to anticipate the correct output from a function, and therefore easier to write sensible tests which help with the process of debugging and validating code.

Initially, the plan was to produce a synthetic dataset to contain abstract and more clearly synthetic data, using simple names and codes for drugs, similar to the data included in the AdhereR package.(241) This was changed in favour of a dataset which uses the format of real Scottish prescribing data, including real drug names, BNF codes, formulations and strengths with a view to being able to produce more detailed and varied examples of the possible applications of the functions on real world data.

The synthetic prescribing dataset produced, “synth_presc”, is based on the characteristics of a typical PIS dataset provided for patients followed up as part of a study. Fields were chosen based on the data required by the functions detailed in this chapter and structured in the same way as they are in PIS data (as described in the Information Services Division [ISD] Scotland PIS fields document, an excerpt of which is adapted in Table 6, below).(242) The generated dataset contains synthetic records for 1,478 individual prescriptions split across 100 patients, with each record containing 13 fields. There are four different drugs in the dataset - simvastatin, atorvastatin, omeprazole and citalopram which are commonly prescribed drugs, meaning there was more data and more variation to draw from when sampling to generate synthetic patients.

Drugs were chosen to represent a number of different drug classes, but two lipid-regulating drugs were chosen to allow for examples showing how results can differ if examining exposure to individual drugs or across drug categories. All of the drugs chosen are most often prescribed as either tablets or capsules and are most often prescribed as 1 unit per day - this was done to avoid overcomplicating the dataset with too wide a range of formulations or dosage instructions.

Field name	Description	Format	Example
patient_id	a unique patient identifier	string (or numeric)	10001
presc_date	the date the prescription was written	date	05/07/2020
approved_name	the approved name of the prescribed item	string	SIMVASTATIN
qty_dispensed	the quantity of the drug prescribed (e.g. number of tablets, volume of liquid)	numeric	28
item_strength	the item strength and the unit of measurement	string	40 MG
bnf_chapter bnf_section bnf_subsection bnf_paragraph	codes corresponding to the location of the prescribed item in the British National Formulary (BNF)	string or numeric	2 212 21200 212000
bnf_item_code	a 15-digit code - the first seven digits detail the BNF categories, and the last eight digits represent the medicinal product form, strength and generic equivalent	string	0212000Y0AAADAD
ddd_conversion	a factor by which the quantity of drug prescribed should be divided to give the number of Defined Daily Doses (DDDs) prescribed	numeric	0.75
ddd_dispensed	the number of DDDs dispensed, based on dispensed quantity and item strength	numeric	37.333
qty_per_day	the prescriber’s instructions for how many units (e.g. tablets, capsules) should be taken per day	numeric	1

Table 6 – An overview of the structure of the subset of PIS data fields used to create the synthetic dataset, adapted from the ISD Data Dictionary for PIS data(242)

In order to determine which patients in the synthetic dataset would be prescribed which drugs each drug of interest was assigned a code within the sample PIS data, as shown in Table 7 below. This was used to create lists of pseudonymised patients who had at least one prescription for each drug. The lists and codes were combined, resulting in a list of patient IDs and codes corresponding to how many of the four drugs they had been prescribed. For example, a patient with at least one prescription for each of the four drugs would be coded ‘1111’, whereas a patient who had only been prescribed omeprazole and simvastatin would be coded ‘1001’. A frequency table for the codes was constructed, and a weighted sample was taken to give drug combinations to assign to the 100 synthetic patients. Each patient was assigned a start date for each of their drugs based on random sampling between 01/01/2020 and 01/01/2021.

Drug	Code
Simvastatin	0001
Atorvastatin	0010
Citalopram	0100
Omeprazole	1000

Table 7 - Codes assigned to the drugs of interest when assessing the frequency of co-prescribing in the real dataset to assign drug combinations to synthetic patients

Then, for each drug, frequency tables were constructed detailing the number of prescriptions per patient, the number of days between successive prescriptions and the strength and formulation prescribed. From these, each patient was assigned a number of prescriptions for each drug they were given, and a corresponding set of differences, in days, between individual prescriptions which were added to the start date successively to give the remaining prescription dates. Patients were also assigned a strength for each drug and a number of tablets per prescription, again determined through weighted sampling and the same for each prescription the patient had for a particular drug. The DDD conversion factor and number of DDDs dispensed were calculated based on the WHO’s DDD index.(243) Every prescription was assigned a tablets per day value of 1 - this means that the PDDs dispensed will differ from the DDDs dispensed in some cases and therefore allows for demonstration of how these measurements differ. These records were then combined to give the complete dataset.

Although the synthetic dataset is based on data characteristics for subjects within a real study, none of the patients represented within it or the individual prescriptions are the same as any of the data within the original study. A second synthetic dataset, “synth_events”, was created to be used alongside the prescribing data in the functions which make use of event dates. This dataset contains the same 100 patient IDs, dates for the start and end of follow-up and up to two event dates per patient. The start date for each patient is the date of their first prescription for any drug, and the end date is the last day of the year in which they had their last prescription. Patients were randomly assigned either 0, 1 or 2 events, and event dates were determined by adding a randomly assigned number of days between 1 and 270 to the start date (for event 1) or the first event date (for event 2) where applicable. Event dates generated after the end of follow-up were removed. A section of this data can be found in Table 8 below.

patient_id	start_date	end_date	event_1	event_2
10001	2020-07-05	2022-12-31	NA	NA
10002	2020-05-29	2022-12-31	2020-06-30	2020-08-06
10003	2020-10-21	2021-12-31	2021-01-22	2021-10-02
10004	2020-09-22	2021-12-31	2021-05-06	NA
10005	2020-08-11	2020-12-31	NA	NA
10006	2020-05-27	2022-12-31	2020-09-06	2021-03-01

Table 8 - Extract from the synthetic events dataset (‘synth_events’) showing the structure of the first 6 records

It is important to note that while these synthetic datasets were generated based on characteristics of real data and can be used to demonstrate different methods of quantifying drug exposure, the results may not always make clinical sense due to the way the data was constructed based on random sampling of a limited number of characteristics from the real data and the fact that each patient was assigned one strength or formulation of a drug, and always received the same number of units per prescription. In particular, for patients with multiple drugs the prescriptions for each drug were generated separately, and so there may be overlap in treatments which do not reflect how the drugs would be prescribed in real clinical settings. Additionally, the event dates and number of events per patient were assigned at random, so links between, for example, adherence or persistence to treatment and time to event will not reflect actual clinical data. The synthetic dataset was to allow for quick validation of newly written or updated code during the development of the package, but all of the functions were also tested on real-world datasets in order to ensure they could handle more complex data.

3.3 Package overview

Table 9 contains a full list of the functions contained within the prescribeR package, as well as the R commands for each function and a short summary of the purpose of each function.

Function name	R command	Description
Data standardisation	tidy_presc	Converts the field names and formats of the input dataset to the prescribeR package standards
Data summaries		
Summarise prescribing dataset	presc_data_summary	Generates a high-level summary of a prescribing dataset
Most commonly prescribed drugs	presc_top_drugs	Determines the most commonly prescribed drugs
Prescription time trends	presc_by_time	Generates a summary of the number of prescriptions over time
Ever use	ever_use	Classified patients as exposed or unexposed based on a threshold number of prescriptions
Use at time point		
Use within a fixed date range	uat_fixed	Determines exposure within a fixed time period for all patients
Use within a fixed range from individual event dates	uat_fixed_events	Determines exposure within a set number of days from a patient-specific event
Use within individual patient date ranges	uat_var_events	Determines exposure between two patient-specific dates
New Users – fixed start date	new_users_fixed	Classifies patients as new or existing users of a drug based on a fixed start date
New users - individual start dates	new_users_var	Classifies patients as new or existing users of a drug based on a patient-specific start date
Current vs. past use at event date	uat_recent	Classifies exposed patients as current or past users of a drug of interest at an event date

Function name	R command	Description
Two prescriptions within a specific timeframe	uat_gap	Determines exposure based on a definition of two prescriptions within a set number of days
Split follow-up into exposure windows	uat_windows	Splits the follow-up period into equal length time periods and determines exposure in each
Daily-dose related methods		
Calculate cumulative daily doses	dd_sum	Calculates the total daily doses prescribed to each patient
Calculate prescription durations	dd_duration	Calculates the expected duration for each prescription based on daily doses prescribed
Calculate prescribed daily doses dispensed	calculate_pdd	Calculated the prescribed daily dose per prescription based on dosage instructions
Persistence		
Refill gap only	refill_gap	Determines persistence to medications based on the number of days between prescriptions only
Refill gap with coverage	refill_gap_dd	Determines persistence to medications based on the number of days between the end of estimated supply and the next prescription

Table 9 - Summary of the functions contained in the prescribeR package

3.4 Data standardisation (tidy_presc)

3.4.1 Rationale

The function described in this section converts the field names and data formats of the user's data to the standards used with the rest of the functions in the prescribeR package. It was created to help maintain consistent object, variable and argument names throughout the rest of the package code to make it easier to read and understand, to provide a standard data structure for use when adding additional functions and to help minimise errors.

The function was initially written as an independent function which the user needed to apply to their data manually before using the other functions in the package, which would require either overwriting the original copy of the data or creating a new copy with the correct formatting. During the process of testing the package, it was decided that neither of these options was ideal. Overwriting the original dataset would mean that if, for example, the user wished to use a different drug identifier they would need to either re-import the dataset or manually undo the changes to field names and formats to run the standardisation function(s) again. Holding multiple copies of a dataset within the global R environment can also be impractical, as it can become difficult to keep track of the different datasets and holding large amounts of data in memory can cause performance issues when running code, particularly when using larger data sets.

In order to make the package more user friendly, the decision was made to have each of the other functions in the package run the standardisation function first. The arguments to the standardisation function correspond to the required fields from the rest of the package and are used to indicate which fields are of interest and to create a new, correctly formatted copy of the data internally before performing any other data transformations. This allows users to try different analysis options with relative ease, and without the need to make manual changes to their original datasets. For example, the 'drug_id' field is used throughout to filter the data for prescriptions for the drug(s) of interest. By selecting the relevant field within their prescribing data, the user can filter by different identifiers such as approved names, brand names or codes such as BNF codes, Anatomical Therapeutic Chemical (ATC) classification system codes or National Drug Codes as desired.

The standardisation function is still available for users to apply directly to their data if they prefer. This allows for the creation of tidy versions of their data within the global R environment, which would cut down on the number of arguments that need to be passed to the other functions as the standard name for each column is the default value for each of the column name arguments. Table 10 below gives the details of the fields used by the other functions, as well as the standard names and formats these fields are converted to by the standardisation function.

	Field name	Data type	Example(s)
Patient ID	patient_id	character	10001
Drug ID	drug_id	character	SIMVASTATIN 212000
Prescription date	presc_date_x	date	2020-01-31
Quantity dispensed	qty_disp	number	56
DDs dispensed	dd_disp	number	28
Quantity per day	qty_per_day	number	1
Event date(s)	ev_date_1 ev_date_2	date	2021-12-31

Table 10 - Overview of the standardised field names and data types used throughout the prescribeR package

3.4.2 Description

This function standardises the field names and formats of a chosen data frame (df). The user specifies the current name of each desired field to the matching argument, and the function returns a copy of the data frame with the standard field names. Each of the column name arguments is optional, and the function will check if they have been entered before attempting to rename the relevant column. Once a column has been renamed, the function converts its contents to the standard formats as described in Table 10. The user needs also to provide the format of the dates contained in their data to the ‘date_format’ argument as a string containing an R date format (e.g. for dates stored as “2000/12/31” the format is “%Y/%m/%d”). If defined incorrectly, R will coerce the data contained in the relevant fields to a date object using the format provided, which can result in incorrect dates. For example, if the date “31/12/2012” is passed to R as a string for conversion and the format argument provided is

“%Y/%m/%d”, R will convert the dates incorrectly and return “0031-12-20”. all of the transformations have been completed, the function outputs the reformatted copy of the data. Additionally, all of the dates in the chosen field need to be in the stated format to prevent erroneous conversions. Each of the other functions in the prescribeR has arguments for each of the required data fields which are passed to the ‘tidy_presc’ function to standardise the input data before analysis – the same function is used for both prescribing and events data in functions which use both, with a separate call to the function used for each dataset.

3.4.3 Arguments

- df - a data frame containing prescribing records to be standardised
- patient_id_col - a string, the name of the column in df containing the patient IDs
- drug_id_col - a string, the name of the column in df containing the drug identifier to be used
- presc_date_col - a string, the name of the column in df containing the prescription date
- dd_disp_col - a string, the name of the column in df containing the number of daily doses of the drug dispensed
- qty_disp_col - a string, the name of the column containing the quantity of the drug dispensed
- qty_per_day_col - a string, the name of the column containing the quantity of drug to be taken per day
- ev_date_1_col - a string, the name of the column containing the first event date

- ev_date_2_col - a string, the name of the column containing the second set of event dates

- date_format - a string containing the format of the dates used in df

3.4.4 R Code

```r
tidy_presc <- function(df,
                       patient_id_col = NULL,
                       drug_id_col = NULL,
                       presc_date_col = NULL,
                       dd_disp_col = NULL,
                       qty_disp_col = NULL,
                       qty_per_day_col = NULL,
                       ev_date_1_col = NULL,
                       ev_date_2_col = NULL,
                       date_format) {
  df1 <- df
  if (!is.null(patient_id_col)) {
    df1 <- df1 %>%
      dplyr::rename(patient_id = patient_id_col)
    df1$patient_id <- as.character(df1$patient_id)
  }
  if (!is.null(drug_id_col)) {
    df1 <- df1 %>%
      dplyr::rename(drug_id = drug_id_col)
    df1$drug_id <- as.character(df1$drug_id)
  }
  if (!is.null(presc_date_col)) {
    df1 <- df1 %>%
      dplyr::rename(presc_date_x = presc_date_col)
    df1$presc_date_x <-
      as.Date(df1$presc_date_x, format = date_format)
  }
  if (!is.null(dd_disp_col)) {
    df1 <- df1 %>%
      dplyr::rename(dd_disp = dd_disp_col)
    df1$dd_disp <- as.numeric(df1$dd_disp)
  }
  if (!is.null(qty_disp_col)) {
    df1 <- df1 %>%
      dplyr::rename(qty_disp = qty_disp_col)
    df1$qty_disp <- as.numeric(df1$qty_disp)
  }
  if (!is.null(qty_per_day_col)) {
    df1 <- df1 %>%
      dplyr::rename(qty_per_day = qty_per_day_col)
    df1$qty_per_day <- as.numeric(df1$qty_per_day)
  }
  if (!is.null(ev_date_1_col)) {
    df1 <- df1 %>%
      dplyr::rename(ev_date_1 = ev_date_1_col)
    df1$ev_date_1 <- as.Date(df1$ev_date_1, format = date_format)
  }
  if (!is.null(ev_date_2_col)) {
    df1 <- df1 %>%
      dplyr::rename(ev_date_2 = ev_date_2_col)
    df1$ev_date_2 <- as.Date(df1$ev_date_2, format = date_format)
  }
  return(df1)
}
```

3.4.5 Usage example

In the example below, the input data contains a single prescription record, with a similar structure to the ‘synth_presc’ data and field names and formats which are not those used by the prescribeR functions. The names of the columns of interest are passed to the corresponding arguments, and the function returns a modified version of the input data frame. Only the specified columns are altered - in the output shown below, for example, the ‘presc_date’ field has been converted to the date format and relabelled and the ‘drug_name’ field has been relabelled as ‘drug_id’, whereas the ‘disp_date’ field has not been changed.(235)

Input:

pat_id	presc_date	disp_date	drug_name	bnf_chapter
<int>	<chr>	<chr>	<chr>	<chr>
10001	12/06/2020	18/06/2020	SIMVASTATIN	02

```
synth_presc_tidy <- tidy_presc(
  df = synth_presc,
  patient_id_col = "patient_id",
  drug_id_col = "approved_name",
  presc_date_col = "presc_date",
  dd_disp_col = "ddd_dispensed",
  date_format = "%Y-%m-%d"
)
```

Output:

patient_id	presc_date_x	disp_date	drug_id	bnf_chapter
<chr>	<date>	<chr>	<chr>	<chr>
10001	2020-06-12	18/06/2020	SIMVASTATIN	02

3.5 Functions for deriving data summaries

These functions all create tables containing descriptive statistics based on the chosen prescribing dataset. By default, they provide summaries of the whole dataset, but the ‘drug’ argument gives users the option to limit the results to specific drug IDs - this can mean specific drug names, classes such as BNF categories, or some other grouping depending on which field is provided to the drug ID column argument.

3.5.1 Summarise prescribing data (presc_data_summary)

3.5.1.1 Description

This function generates a summary of the prescription data contained within a data frame of interest. The data are tidied by the ‘tidy_presc’ function, then filtered to remove prescriptions where the ‘drug_id’ value does not match the value of the ‘drug’ argument. The number of prescriptions, number of unique drugs (based on distinct values in the ‘drug_id’ column) and the date of the first and last prescription for each patient are calculated. If the ‘summary’ argument is FALSE, this is returned as the output. If the ‘summary’ argument is TRUE, the total number of patients, total number of prescriptions and the date range covered by the data as a whole are calculated, along with the median and interquartile range of the number of prescriptions per patient, and this summary is returned as the output.

3.5.1.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- drug - a string representing drug identifier of interest - if no value is entered, the counts and dates returned represent the dataset as a whole. This argument accepts regular expressions to allow for searching based on exact codes or sections of codes corresponding to e.g. drug groups

- summary – logical – if FALSE the function returns the full results for all patients and if true the function returns summary values describing the dataset as a whole; set to TRUE by default

- patient_id_col, drug_id_col, presc_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.5.1.3 R Code

```
presc_data_summary <-
  function(df,
           drug = ".",
           summary = TRUE,
           patient_id_col = "patient_id",
           drug_id_col = "drug_id",
           presc_date_col = "presc_date_x",
           date_format) {
    tidy_df <-
      tidy_presc(
        df,
        patient_id_col = patient_id_col,
        drug_id_col = drug_id_col,
        presc_date_col = presc_date_col,
        date_format = date_format
      )
    summ1 <- tidy_df %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    summ1 <- summ1 %>%
      dplyr::group_by(.data$patient_id) %>%
      dplyr::summarise(
        n_presc = dplyr::n(),
        n_drugs = dplyr::n_distinct(.data$drug_id),
        first_presc = min(.data$presc_date_x),
        last_presc = max(.data$presc_date_x)
      )
    if (summary == TRUE) {
      summ1 <- summ1 %>%
        dplyr::summarise(
          total_presc = sum(.data$n_presc),
          n_pat = dplyr::n_distinct(.data$patient_id),
          median_n_presc = stats::median(.data$n_presc),
          iqr_n_presc = stats::IQR(.data$n_presc),
          first_presc = min(.data$first_presc),
          last_presc = max(.data$last_presc)
        )
    }
    return(summ1)
  }
```

3.5.1.4 Usage examples

The two examples below provide summaries of the ‘synth_presc’ dataset, highlighting the different outputs which can be obtained from this function based on the value of the ‘summary’ argument. In the first example, the summary argument is set to TRUE, so the results are returned for each patient individually. In the second example, the ‘summary’ argument is FALSE, so the function performs the extra step of summarising the individual patient results before providing output.

3.5.1.4.1 Example 1

```
ex1 <- presc_data_summary(df = synth_presc,
                          summary = TRUE,
                          patient_id_col = "patient_id",
                          drug_id_col = "approved_name",
                          presc_date_col = "presc_date")
```

Output:

patient_id	n_presc	n_drugs	first_presc	last_presc
10001	12	1	2020-07-05	2022-03-01
10002	13	1	2020-05-29	2022-02-20
10003	6	2	2020-10-21	2021-05-22
10004	8	1	2020-09-22	2021-08-02
10005	5	1	2020-08-11	2020-11-27
10006	15	3	2020-05-27	2022-01-01

3.5.1.4.2 Example 2

```
ex2 <- presc_data_summary(df = synth_presc,
                          summary = FALSE,
                          patient_id_col = "patient_id",
                          drug_id_col = "approved_name",
                          presc_date_col = "presc_date")
```

Output:

total_presc	n_pat	median_n_presc	iqr_n_presc	first_presc	last_presc
1478	100	13	10.75	2020-01-02	2023-09-17

3.5.2 Most commonly prescribed drugs (presc_top_drugs)

3.5.2.1 Description

This function generates a list of the most commonly prescribed drugs within the dataset (based on unique values in the drug ID field). The chosen data frame is filtered by the drug identifier provided to the ‘drug’ argument, and the prescriptions for each ID within the remaining data are counted. By default, the top 10 drugs and corresponding counts are returned, but the number of drugs returned can be changed by adjusting the ‘rank’ argument. The function returns a data frame containing the top drug identifiers and the number of prescriptions in the database for the requested number of drugs, in descending order from the most commonly prescribed.

3.5.2.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a prescription date and a drug identifier

- drug - a string containing a drug identifier of interest - if no value is entered the function uses the whole data set. This argument accepts regular expressions

- rank - the number of drugs to return; default is to return the top 10 drugs

- drug_id_col - a string, the name of the column in df containing the drug_id variable, passed to the tidy_presc function

3.5.2.3 R Code

```
presc_top_drugs <-
  function(df,
           drug = "*",
           rank = 10,
           drug_id_col = "drug_id") {
    tidy_df <- tidy_presc(df, drug_id_col = drug_id_col)
    freq <- tidy_df %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    freq <- freq %>%
      dplyr::group_by(.data$drug_id) %>%
      dplyr::summarise(n_presc = dplyr::n())
    freq <- freq %>%
      dplyr::top_n(rank, .data$n_presc) %>%
      dplyr::arrange(dplyr::desc(.data$n_presc))
    return(freq)
  }
```

3.5.2.4 Usage example

In the example shown below the output from the function is a list of the top 3 most commonly prescribed drugs in the ‘synth_presc’ dataset by approved drug name (as defined by the ‘drug_id_col’ argument. The number of entries corresponds to the value specified to the ‘rank’ argument, with the number of prescriptions for each value calculated.

```
ex1 <- presc_top_drugs(df = synth_presc,
                       rank = 3,
                       drug_id_col = "approved_name")
```

drug_id	n_presc
SIMVASTATIN	797
OMEPRAZOLE	416
ATORVASTATIN	183

3.5.3 Prescription time trends (presc_by_time)

3.5.3.1 Description

This function creates a breakdown of the number of prescriptions matching a chosen drug identifier over time. The dataset is filtered by the drug ID provided to the ‘drug’ argument if applicable. The total number of daily doses of each drug dispensed per date to each patient is calculated. If the input data does not contain the number of daily doses dispensed, a dummy variable is generated and then removed before the output is generated. If the ‘flatten’ argument is set to TRUE, the function reduces the dataset so that it contains only one prescription per drug ID and patient for each date. The month, year, quarter and semester are then extracted from each prescription’s date - quarters and semesters are defined by splitting the year into three and six month periods respectively. The prescriptions are grouped by one of these values depending on the value of the ‘group’ argument, and the number of prescriptions and number of daily doses dispensed in each group is calculated. Output for this function is a data frame containing number of prescriptions, within each date group, along with the number of DDs dispensed if present in the data

3.5.3.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- drug - a string corresponding to a drug identifier to be matched, regular expressions allowed

- flatten -logical, if TRUE the function only counts one prescription per patient for each drug per date; default value is FALSE

- group - a string corresponding to the desired grouping for results - either month ("M"), quarter ("Q"), semester ("S") or year ("Y"); default is to group by year

- patient_id_col, drug_id_col, presc_date_col, dd_disp_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.5.3.3 R Code

```
presc_by_time <- function(df,
                          drug = "*",
                          flatten = FALSE,
                          group = "Y",
                          drug_id_col = "drug_id",
                          patient_id_col = "patient_id",
                          presc_date_col = "presc_date_x",
                          dd_disp_col = NULL,
                          date_format) {
  tidy_df <-
    tidy_presc(
      df,
      drug_id_col = drug_id_col,
      patient_id_col = patient_id_col,
      presc_date_col = presc_date_col,
      dd_disp_col = dd_disp_col,
      date_format = date_format
    )
  tidy_df <- tidy_df %>%
    dplyr::filter(grepl(drug, .data$drug_id))
  if (is.null(dd_disp_col)) {
    tidy_df <- tidy_df %>%
      dplyr::mutate(dd_disp = 0)
  }
  if (flatten == TRUE) {
    tidy_df <- tidy_df %>%
      dplyr::group_by(.data$patient_id, .data$presc_date_x, .data$drug_i
d) %>%
      dplyr::mutate(sum_dd = sum(.data$dd_disp)) %>%
      dplyr::ungroup() %>%
      dplyr::select(.data$patient_id,
                    .data$presc_date_x,
                    .data$drug_id,
                    dd_disp = .data$sum_dd) %>%
      dplyr::distinct()
  }
  summ1 <- tidy_df %>%
    dplyr::mutate(
      presc_month = lubridate::month(.data$presc_date_x),
      presc_year = lubridate::year(.data$presc_date_x),
      presc_quarter = lubridate::quarter(.data$presc_date_x, with_year =
TRUE),
      presc_semester = lubridate::semester(.data$presc_date_x, with_year
= TRUE)
    ) %>%
    dplyr::select(
      .data$patient_id,
      .data$presc_date_x,
      .data$drug_id,
      .data$dd_disp,
      .data$presc_month,
      .data$presc_year,
      .data$presc_quarter,
      .data$presc_semester
    )
  if (group == "Y") {
    result <- summ1 %>%
      dplyr::group_by(.data$presc_year) %>%
      dplyr::summarise(
        n_presc = dplyr::n(),
        n_patients = dplyr::n_distinct(.data$patient_id),
        total_dds = sum(.data$dd_disp)
      )
  } else if (group == "M") {
    result <- summ1 %>%
      dplyr::group_by(.data$presc_year, .data$presc_month) %>%
      dplyr::summarise(
        n_presc = dplyr::n(),
        n_patients = dplyr::n_distinct(.data$patient_id),
        total_dds = sum(.data$dd_disp)
      )
  } else if (group == "S") {
    result <- summ1 %>%
      dplyr::group_by(.data$presc_semester) %>%
      dplyr::summarise(
        n_presc = dplyr::n(),
        n_patients = dplyr::n_distinct(.data$patient_id),
        total_dds = sum(.data$dd_disp)
      )
  } else if (group == "Q") {
    result <- summ1 %>%
      dplyr::group_by(.data$presc_quarter) %>%
      dplyr::summarise(
        n_presc = dplyr::n(),
        n_patients = dplyr::n_distinct(.data$patient_id),
        total_dds = sum(.data$dd_disp)
      )
  }
  if (is.null(dd_disp_col)) {
    result <- result %>%
      dplyr::select(-.data$total_dds)
  }
  return(result)
}
```

3.5.3.4 Usage examples

The examples shown below highlight different outputs available using this function. In the first example, the number of prescriptions and number of patients with at least 1 prescription for atorvastatin are returned by year. Additionally, the ‘flatten’ argument is set to TRUE, meaning the total number of prescriptions returned only counts one prescription per patient per date. In the second example, the results are returned by quarter instead of by year. The output also contains the total DDDs dispensed as a column name was passed to the ‘dd_disp_col’ argument. The ‘drug_id’ column of interest is the BNF section, so the results represent the number of patients prescribed any drug in that section each quarter.

3.5.3.4.1 Example 1

```
ex1 <- presc_by_time(df = synth_presc,
                     drug = "ATORVASTATIN",
                     group = "Y",
                     flatten = TRUE,
                     drug_id_col = "approved_name",
                     presc_date_col = "presc_date")
```

Output:

presc_year	n_presc	n_patients
2020	94	23
2021	68	14
2022	16	5
2023	5	2

3.5.3.4.2 Example 2

```
ex2 <- presc_by_time(df = synth_presc,
                     drug = "212",
                     group = "Q",
                     drug_id_col = "bnf_section",
                     presc_date_col = "presc_date",
                     dd_disp_col = "ddd_dispensed")
```

```
Output:
presc_quarter  n_presc  n_patients  total_dds
2020.1         22       14          1260.000
2020.2         69       39          4559.333
2020.3         113      62          6757.333
2020.4         151      79          8549.333
2021.1         150      78          8885.333
2021.2         134      71          7728.000
```

3.6 Ever use (ever_use)

3.6.1 Rationale

The simplest method for defining drug exposure based on prescribing data is to classify any subject who has ever had a prescription for the drug of interest as exposed, and those who have not as unexposed. Splitting subjects into two groups in this manner allows for the comparison of outcomes between those who are unexposed and exposed to assess the effectiveness of the drug of interest. Additionally, determining the subjects who were exposed to the drug of interest allows for further analysis of these subjects, for example assessing different levels of exposure or patterns of utilisation of the drug of interest. For the most part, ever use provides a dichotomous variable, which can be useful as a factor for inclusion in statistical models of risk.

The studies reviewed in the previous chapter highlighted some variants and extensions to the ever use methodology. The threshold for classification of exposure could be higher than just one prescription in order to try and focus in on recurrent use as opposed to incident use, or cases where the drug was discontinued soon after prescription. In cases where a higher threshold is used, patients who have been prescribed the drug but did not meet the threshold for exposure are classified as unexposed. The number of prescriptions may also be of use in some cases to further classify extent of exposure, providing a continuous variable in addition to the categorical classification of exposure. After determining ever use, the date of the first prescription is often then used as an index date for exposure to determine when the patients entered into the cohort being created. Additionally, as well as considering ever use of a single drug, it could also be useful to consider more than one drug as a measure of co-prescribing or polypharmacy, or exposure to one or more of a group of drugs, e.g. all of the drugs appearing with a particular section of the BNF, as opposed to a single drug of interest.

3.6.2 Description

The function takes a data frame containing records of individual prescriptions (df) and applies the ‘tidy_presc’ function using the column IDs provided. A list of all of the patient IDs contained within df is created. The data are then filtered to remove prescriptions which do not match the drug identifier provided in the ‘drug’ argument. If the ‘flatten’ argument is set to true, the function filters the dataset so that only one prescription per unique drug identifier is included for each patient per date. Next, the rows are grouped by patient ID and the date of the first prescription and number of prescriptions for each is determined. The function then assigns a flag to each patient record indicating if they meet the threshold for exposure (1) or not (0). In cases where the value of the ‘threshold’ argument is not 1, patients who have prescriptions for the drug(s) but do not meet the threshold are classified as unexposed. If the ‘return_all’ argument is set to FALSE, patients who did not meet the threshold for exposure are removed from the result. If the ‘return_all’ argument Is set to TRUE, the result contains all of the patients from the input along with the exposure flags. If the ‘summary’ argument is set to FALSE, only the patient IDs and exposure flags are returned, whereas if the ‘summary’ variable is set to TRUE, the dates of first prescriptions and the total number of prescriptions per patient are also included in the output.

3.6.3 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- drug – A string containing the drug identifier be matched, regular expressions allowed

- flatten – logical, if TRUE the function only counts one prescription per unique drug identifier per date for each patient, default value is FALSE

- threshold – The minimum number of prescriptions that must be present for the patient’s ID to be included in the results; default value is 1
- summary – If TRUE, the results returned will include the total number of prescriptions and the date of the first prescription for each patient ID where the number of prescriptions was more than the threshold value, default value is FALSE
- return_all – logical, if TRUE the function returns all patient IDs from the original dataset along with a flag indicating if they met the threshold for exposure or not. If FALSE, the function only returns the details of the patients who met the threshold, default value is FALSE
- patient_id_col, drug_id_col, presc_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function
- date_format - a string containing the format of the dates used in df
-

3.6.4 R code

```
ever_use <-
  function(df,
           drug,
           flatten = FALSE,
           threshold = 1,
           summary = FALSE,
           return_all = FALSE,
           patient_id_col = "patient_id",
           drug_id_col = "drug_id",
           presc_date_col = "presc_date_x",
           date_format) {
    tidy_df <-
      tidy_presc(
        df,
        patient_id_col = patient_id_col,
        drug_id_col = drug_id_col,
        presc_date_col = presc_date_col,
        date_format = date_format
      )
    ids <- tidy_df %>%
      dplyr::select(.data$patient_id) %>%
      dplyr::distinct()
    ever1 <- tidy_df %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    if (flatten == TRUE) {
      ever1 <- ever1 %>%
        dplyr::distinct(.data$patient_id, .data$drug_id, .data$presc_dat
e_x)
    }
    ever1 <- ever1 %>%
      dplyr::group_by(.data$patient_id) %>%
      dplyr::summarise(n_presc = dplyr::n(),
                       first_presc = min(.data$presc_date_x)) %>%
      dplyr::mutate(exposed = ifelse(.data$n_presc >= threshold, 1, 0))
    if (return_all == FALSE) {
      ever1 <- ever1 %>%
        dplyr::filter(.data$exposed == 1)
    } else if (return_all == TRUE) {
      ever1 <- dplyr::left_join(ids, ever1, by = "patient_id")
      ever1$n_presc[is.na(ever1$n_presc)] <- 0
      ever1$exposed[is.na(ever1$exposed)] <- 0
    }
    if (summary == TRUE) {
      return(ever1)
    } else if (summary == FALSE) {
      ever1 <- ever1 %>%
        dplyr::select(.data$patient_id, .data$exposed)
      return(ever1)
    }
  }
```

3.6.5 Usage examples

In the first example below, the output contains a list of IDs corresponding to patients who had at least 1 prescription for omeprazole in the ‘synth_presc’ dataset. Since the summary and ‘return_all’ arguments are both FALSE by default, only the IDs for patients who met the threshold were returned. In the second example, the ‘summary’ argument is TRUE, so the number of prescriptions and the date of the first prescription are also returned for each patient. As the ‘return_all’ argument is set to TRUE, data was returned for all patients, with the flag indicating if they were exposed (=1) or unexposed (=0) based on the provided threshold. In this case, the ‘threshold’ argument was set to 2, meaning that patients with only 1 prescription are classified as unexposed.

3.6.5.1.1 Example 1

```
ex1 <- ever_use(df = synth_presc,
                drug = "OMEPRAZOLE",
                summary = FALSE,
                patient_id_col = "patient_id",
                drug_id_col = "approved_name",
                presc_date_col = "presc_date")
```

Output

patient_id	exposed
10003	1
10006	1
10008	1
10009	1
10010	1
10012	1

3.6.5.1.2 Example 2

```
ex2 <- ever_use(df = synth_presc,
                drug = "OMEPRAZOLE",
                summary = TRUE,
                flatten = TRUE,
                threshold = 2,
                return_all = TRUE,
                patient_id_col = "patient_id",
                drug_id_col = "approved_name",
                presc_date_col = "presc_date")
```

Output

patient_id	n_presc	first_presc	exposed
10001	0	NA	0
10002	0	NA	0
10003	2	2020-10-21	1
10004	0	NA	0
10005	0	NA	0
10006	1	2020-12-10	0

3.7 Use at time point

3.7.1 Rationale

Use at time point methods build on the ever use method described above by considering whether or not patients were prescribed a drug during a specific time period of interest. As with ever use, this often involves generating a categorical exposure variable, and can be a way of selecting a cohort of patients for further analysis - the definition of the timeframe of interest will depend on the research question. For example, exposure before a date of interest, either at any time or within a set number of days, could be considered as part of assessing the impact of exposure to the drug of interest on the risk of an outcome of interest such as hospitalisation or death. This can be further refined by splitting subjects into those who were current or past users at the time of an event based on how recent their last prescription before the event date was. Exposure after a date of interest, including at set time points or within specific intervals can be used when examining effectiveness ongoing drug therapy after an event such as hospitalisation. Exposure can also be defined as a time-varying covariate by splitting follow-up time into windows of equal length and then determining whether patients were prescribed the drug of interest during each window.

3.7.2 Fixed date range (uat_fixed)

3.7.2.1 Description

This function determines which patients were exposed to the drug(s) of interest within a standardised timeframe for all patients. After tidying the data through a call to the 'tidy_presc' function, the dataset is filtered to select prescriptions for the drug(s) of interest based on the drug identifier provided ('drug'). If the 'flatten' argument is set to true, the data are filtered to include only one prescription per unique drug ID for each patient per date. The definition of the timeframe of interest is then determined based on the 'date_1', 'timeframe' and 'forward' arguments. If the 'timeframe' argument is set to 0, the function filters for all prescriptions either before ('forward = FALSE') or after ('forward = TRUE') the date of interest entered in 'date_1'. For other values of the 'timeframe' argument, the function determines a second date by either adding or subtracting that number of days, based on the 'forward' variable as above, and then filters for prescription dates that fall within this timeframe. The function then returns the number of prescriptions and first prescription date within the timeframe of interest for each patient.

3.7.2.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- drug - a string containing an identifier for the drug of interest, regular expressions allowed

- flatten - logical, if TRUE the function only counts one prescription per unique drug identifier per date for each patient

- date_1 - a string containing the date of interest - depending on the value of forward, this is either the start or end date of the time window of interest

- timeframe - the number of days in the time frame of interest - defaults to 0, which means the function will check for any prescriptions before or after the date of interest depending on the value of forward

- forward - when TRUE, the function will check for prescriptions after the date of interest, or when a value for timeframe is provided it will be added to the date to determine the end date. When FALSE, function checks for prescriptions before the date of interest, or the value of timeframe is subtracted to determine the start date of the window; set to TRUE by default

- patient_id_col, drug_id_col, presc_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.7.2.3 R code

```
uat_fixed <-
  function(df,
           drug,
           date_1,
           flatten = FALSE,
           timeframe = 0,
           forward = TRUE,
           patient_id_col = "patient_id",
           drug_id_col = "drug_id",
           presc_date_col = "presc_date_x",
           date_format) {
    tidy_df <-
```

```
      tidy_presc(
        df,
        patient_id_col = patient_id_col,
        drug_id_col = drug_id_col,
        presc_date_col = presc_date_col,
        date_format = date_format
      )
    uat1 <- tidy_df %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    if (flatten == TRUE) {
      uat1 <- uat1 %>%
        dplyr::distinct(.data$patient_id, .data$drug_id, .data$presc_dat
e_x)
    }
    date_1 <- as.Date(date_1, format = date_format)
    if ((forward == TRUE) && (timeframe == 0)) {
      uat1 <- uat1 %>%
        dplyr::filter(.data$presc_date_x >= date_1)
    } else if ((forward == FALSE) && (timeframe == 0)) {
      uat1 <- uat1 %>%
        dplyr::filter(date_1 >= .data$presc_date_x)
    } else if ((forward == TRUE) && (timeframe > 0)) {
      date_2 <- date_1 + timeframe
      uat1 <- uat1 %>%
        dplyr::filter(.data$presc_date_x >= date_1 &
                        .data$presc_date_x <= date_2)
    } else if ((forward == FALSE) && (timeframe > 0)) {
      date_2 <- date_1 - timeframe
      uat1 <- uat1 %>%
        dplyr::filter(.data$presc_date_x <= date_1 &
                        .data$presc_date_x >= date_2)
    }
    uat_result <- uat1 %>%
      dplyr::group_by(.data$patient_id) %>%
      dplyr::summarise(n_presc = dplyr::n(),
                       first_presc = min(.data$presc_date_x))
    return(uat_result)
  }
```

3.7.2.4 Usage examples

The examples below highlight how the ‘timeframe’ argument alters the output from this function. In the first example, the value of the argument is set to 0, which means that the output contains details of the patients who had at least 1 prescription for citalopram at any time after 01/01/2021, the value passed to the ‘date_1’ argument. In the second example, the ‘timeframe’ argument is set to 180, so only prescriptions in the 180 days after 01/01/2021 are used to define exposure, and therefore the number of prescriptions for some patients is lower than in the first example.

3.7.2.4.1 Example 1

```r
ex1 <- uat_fixed(df = synth_presc,
                 drug = "CITALOPRAM",
                 date_1 = "2021-01-01",
                 timeframe = 0,
                 forward = TRUE,
                 patient_id_col = "patient_id",
                 drug_id_col = "approved_name",
                 presc_date_col = "presc_date")
```

patient_id	n_presc	first_presc
10009	8	2021-01-22
10035	2	2021-02-22
10041	7	2021-01-17
10061	6	2021-01-23
10076	8	2021-01-21
10083	3	2021-01-08

3.7.2.4.2 Example 2

```r
ex2 <- uat_fixed(df = synth_presc,
                 drug = "CITALOPRAM",
                 date_1 = "2021-01-01",
                 timeframe = 180,
                 forward = TRUE,
                 patient_id_col = "patient_id",
                 drug_id_col = "approved_name",
                 presc_date_col = "presc_date")
```

patient_id	n_presc	first_presc
10009	4	2021-01-22
10035	2	2021-02-22
10041	5	2021-01-17
10061	3	2021-01-23
10076	5	2021-01-21
10083	3	2021-01-08

3.7.3 Fixed range from individual patient event dates (uat_fixed_events)

3.7.3.1 Description

This function determines if patients had at least 1 prescription for the drug(s) of interest within a timeframe of interest, based on a standardised timeframe from patient-specific event dates. The prescription (‘df’) and events data (‘df2’) are standardised using the ‘tidy_presc’ function, the event dates contained in ‘df2’ are joined to the prescribing data, matched by the patient IDs and patients without a listed event date are removed. The data are filtered for prescriptions matching the drug ID provided. If the ‘flatten’ argument is set to TRUE, the data are filtered to include only one prescription for each drug ID per date for each patient. If the timeframe argument is set to 0, the function filters for prescriptions at any point before or after before the event date (if ‘forward = FALSE’) or after the event date (‘forward = TRUE’). If the timeframe argument is not equal to zero, the function determines a start or end date by subtracting or adding the timeframe value, in days, from the event date, and filters for prescriptions that fall within this window (again depending on the value of the ‘forward’ argument). The function then determines the first prescription date within the window and the number of prescriptions for each patient and returns these values.

3.7.3.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- df2 – a data frame containing patient event records, consisting of at least a patient ID and an event date

- drug - a string containing the identifier for the drug of interest, regular expressions allowed

- flatten – logical, if TRUE the function only counts one prescription per unique drug identifier per date for each patient, FALSE by default

- timeframe - the number of days in the time frame of interest - defaults to 0, which means the function will check for any prescriptions before or after the date of interest depending on the value of forward

- forward - when TRUE, the function will check for prescriptions after the date of interest, or when a value for timeframe is provided it will be added to the date to determine the end date. When FALSE, function checks for prescriptions before the date of interest, or the value of timeframe is subtracted to determine the start date of the window; set to TRUE by default

- patient_id_col, drug_id_col, presc_date_col, ev_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.7.3.3 R code

```
uat_fixed_events <-
  function(df,
           df2,
           drug,
           flatten = FALSE,
           timeframe = 0,
           forward = TRUE,
           patient_id_col = "patient_id",
           drug_id_col = "drug_id",
           presc_date_col = "presc_date_x",
           ev_date_col = "ev_date_1",
           date_format) {
    tidy_df <-
      tidy_presc(
        df,
        patient_id_col = patient_id_col,
        drug_id_col = drug_id_col,
        presc_date_col = presc_date_col,
        date_format = date_format
      )
    tidy_ev <-
      tidy_presc(
        df2,
        patient_id_col = patient_id_col,
        ev_date_1_col = ev_date_col,
        date_format = date_format
      )
    uat1 <- dplyr::left_join(tidy_df, tidy_ev, by = "patient_id")
    uat1 <- dplyr::filter(uat1,!is.na(.data$ev_date_1))
    uat1 <- uat1 %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    if (flatten == TRUE) {
      uat1 <- uat1 %>%
        dplyr::distinct(.data$patient_id, .data$drug_id, .data$presc_dat
e_x)
    }
    if ((forward == TRUE) && (timeframe == 0)) {
      uat1 <- uat1 %>%
        dplyr::filter(.data$presc_date_x >= .data$ev_date_1)
    } else if ((forward == FALSE) && (timeframe == 0)) {
      uat1 <- uat1 %>%
        dplyr::filter(.data$ev_date_1 >= .data$presc_date_x)
    } else if ((forward == TRUE) && (timeframe > 0)) {
      uat1 <- uat1 %>%
        dplyr::filter(
          .data$presc_date_x >= .data$ev_date_1 &
            .data$presc_date_x <= .data$ev_date_1 + timeframe
        )
    } else if ((forward == FALSE) && (timeframe > 0)) {
      uat1 <- uat1 %>%
        dplyr::filter(
          .data$presc_date_x <= .data$ev_date_1 &
            .data$presc_date_x >= .data$ev_date_1 - timeframe
        )
    }
    uat_result <- uat1 %>%
      dplyr::group_by(.data$patient_id) %>%
      dplyr::summarise(n_presc = dplyr::n(),
                       first_presc = min(.data$presc_date_x))
    return(uat_result)
  }
```

3.7.3.4 Usage example

The example below shows a summary of the number of patients in the ‘synth_presc’ dataset who had at least 1 prescription for any drug listed in BNF sub-section 2.12.00 in the 180 days before their first event from the ‘synth_events’ dataset. Since the forward argument is ‘TRUE’, the value of ‘timeframe’ is subtracted from the value of ‘event_1’ to determine the timeframe of interest.

```
ex1 <- uat_fixed_events(df = synth_presc,
                        df2 = synth_events,
                        drug = "21200",
                        timeframe = 180,
                        forward = FALSE,
                        patient_id_col = "patient_id",
                        drug_id_col = "bnf_subsection",
                        presc_date_col = "presc_date",
                        ev_date_col = "event_1")
```

patient_id	n_presc	first_presc
10002	1	2020-05-29
10003	1	2020-11-12
10004	4	2021-01-25
10006	2	2020-05-27
10010	1	2020-07-02
10011	4	2020-06-08

3.7.4 Individual patient date ranges (uat_var_events)

3.7.4.1 Description

This function determines if patients were exposed to the drug(s) of interest during timeframes based on patient-specific event dates (e.g. start and end of follow up, between two hospitalisations). Both datasets are standardised using the ‘tidy_presc’ function, the event dates are joined on to the prescribing records based on patient IDs and data for any patients who do not have two event dates listed are removed. The data are then filtered to select prescriptions where the drug ID matches the value provided in the ‘drug’ argument and filtered again to select prescriptions which fall between the two event dates, i.e. the prescribed date value is greater than the first event date but less than the second. If the ‘flatten’ argument is set to true, the function filters the dataset so that only one prescription per unique drug identifier is included for each patient per date. Once the relevant prescriptions are selected, the number of prescriptions per patient and date of each patient’s first prescription are determined and returned.

3.7.4.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- df2 – a data frame containing event records consisting of at least a patient ID two event dates per patient

- drug – a string containing an identifier for the drug of interest, regular expressions allowed

- flatten – logical, if TRUE the function only counts one prescription per unique drug identifier per date for each patient, FALSE by default

- patient_id_col, drug_id_col, presc_date_col, ev_date_1_col, ev_date_2_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.7.4.3 R code

```
uat_var_events <- function(df,
                           df2,
                           drug,
                           flatten = FALSE,
                           patient_id_col = "patient_id",
                           drug_id_col = "drug_id",
                           presc_date_col = "presc_date_x",
                           ev_date_1_col = "ev_date_1",
                           ev_date_2_col = "ev_date_2",
                           date_format) {
  tidy_df <-
    tidy_presc(
      df,
      patient_id_col = patient_id_col,
      drug_id_col = drug_id_col,
      presc_date_col = presc_date_col,
      date_format = date_format
    )
  tidy_ev <- tidy_presc(
    df2,
    patient_id_col = patient_id_col,
    ev_date_1_col = ev_date_1_col,
    ev_date_2_col = ev_date_2_col,
    date_format = date_format
  )
  uat1 <- dplyr::left_join(tidy_df, tidy_ev, by = "patient_id")
  uat1 <-
    dplyr::filter(uat1,!is.na(.data$ev_date_1) &
                    !is.na(.data$ev_date_2))
  uat1 <- uat1 %>%
    dplyr::filter(grepl(drug, .data$drug_id))
  if (flatten == TRUE) {
    uat1 <- uat1 %>%
      dplyr::distinct(.data$patient_id, .data$drug_id, .data$presc_date_
x)
  }
    uat1 <- uat1 %>%
      dplyr::filter(.data$presc_date_x > .data$ev_date_1 &
                      .data$presc_date_x < .data$ev_date_2)
  uat_result <- uat1 %>%
    dplyr::group_by(.data$patient_id) %>%
    dplyr::summarise(n_presc = dplyr::n(),
                     first_presc = min(.data$presc_date_x))
  return(uat_result)
}
```

3.7.4.4 Usage example

The example below shows the format of the output from this function, in this case a list of patients who were exposed to a drug from BNF section 4.03 in between two hospitalisations (‘event_1’ and ‘event_2’). The definition of exposure in this function requires a value for both event dates, so only patients who have an event listed in both of the specified fields and at least 1 prescription between those two dates are included in the output.

```
ex1 <- uat_var_events(df = synth_presc,
                      df2 = synth_events,
                      drug = "403",
                      patient_id_col = "patient_id",
                      drug_id_col = "bnf_section",
                      presc_date_col = "presc_date",
                      ev_date_1_col = "event_1",
                      ev_date_2_col = "event_2")
```

patient_id	n_presc	first_presc
10061	1	2020-09-11
10076	1	2020-10-23
10083	2	2020-09-11

3.7.5 New users – fixed start date (new_users_fixed)

3.7.5.1 Description

This function determines whether patients are classified as new or existing users of the drug(s) of interest at a fixed start date. The prescribing records are standardised using the ‘tidy_presc’ function, and a copy of all patient IDs in the data are stored for later use. The records are filtered for prescriptions matching the drug ID provided in the ‘drug’ argument. If the timeframe argument is equal to 0, any patient with a prescription before the date given in the ‘start_date’ argument is flagged as an existing user. If the ‘timeframe’ argument is not equal to zero, the value of ‘timeframe’ is subtracted from ‘start_date’ to give the start of the lookback period, and any patients with prescriptions in that period is flagged as an existing user. Patients are then flagged as exposed or unexposed based on if they have at least 1 prescription during the period after the start date, and the first prescription date is recorded where applicable. The flags for existing users are then combined with the exposure status during follow-up, and any patients who are not flagged as existing users but are flagged as exposed during follow-up are flagged as new users. If the ‘return_all’ argument is FALSE, the data are then filtered to contain only new users exposed during follow-up. Otherwise, data for all patients is returned, with a flag indicating exposure and new user status and the date of the first prescription for the drug(s) of interest after the start date.

3.7.5.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- drug – a string containing an identifier for the drug of interest, regular expressions allowed

- start_date – a string, containing the date to be used as the start of follow-up in the same format as the dates contained in df

- timeframe – a number, the desired length of lookback period in days. If set to 0, all prescriptions before the start date will be used, default value is 0

- return_all – logical, if TRUE only records for new users exposed during the period after the start date will be returned, if FALSE records will be returned for all patients with flags indicating if they were exposed and if they are a new user; default value is TRUE

- patient_id_col, drug_id_col, presc_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.7.5.3 R code

```
new_users_fixed <-
  function(df,
           drug,
           start_date = NULL,
           timeframe = 0,
           return_all = TRUE,
           patient_id_col = "patient_id",
           drug_id_col = "drug_id",
           presc_date_col = "presc_date_x",
           date_format) {
    tidy_df <-
      tidy_presc(
        df,
        patient_id_col = patient_id_col,
        drug_id_col = drug_id_col,
        presc_date_col = presc_date_col,
        date_format = date_format
      )
    ids <- tidy_df %>%
      dplyr::select(.data$patient_id) %>%
      dplyr::distinct()
    tidy_df <- tidy_df %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    start_date <- as.Date(start_date, format = date_format)
    if (timeframe == 0) {
      uat1 <- tidy_df %>%
        dplyr::group_by(.data$patient_id) %>%
        dplyr::filter(.data$presc_date_x < start_date) %>%
        dplyr::distinct(.data$patient_id) %>%
        dplyr::mutate(new_user = 0)
    } else if (timeframe != 0) {
      uat1 <- tidy_df %>%
        dplyr::group_by(.data$patient_id) %>%
        dplyr::filter(.data$presc_date_x < start_date &
                        .data$presc_date_x > start_date - timeframe) %>%
        dplyr::distinct(.data$patient_id) %>%
        dplyr::mutate(new_user = 0)
    }
    uat2 <- tidy_df %>%
      dplyr::group_by(.data$patient_id) %>%
      dplyr::filter(.data$presc_date_x >= start_date) %>%
      dplyr::summarise(first_presc = min(.data$presc_date_x)) %>%
      dplyr::mutate(exposed = 1)
    uat1 <- dplyr::left_join(ids, uat1, by = "patient_id")
    uat1 <- dplyr::left_join(uat1, uat2, by = "patient_id")
    uat1$new_user[is.na(uat1$new_user) & uat1$exposed == 1] <- 1
    uat1$exposed[is.na(uat1$exposed)] <- 0
    if (return_all == FALSE) {
      uat1 <- uat1 %>%
        dplyr::filter(.data$exposed == 1 & .data$new_user == 1)
      return(uat1)
    } else {
      return(uat1)
    }
  }
```

3.7.5.4 Usage example

In the example below, the patients in ‘synth_presc’ are split into new and existing users based on a lookback period of 45 days from 01/01/2021. In the output, patients who have a ‘new_user’ flag value of 1 did not have a prescription for the drug of interest in the lookback period defined, whereas patients who have a value of 0 had at least 1. Patients flagged as exposed had at least 1 prescription for the drug of interest after this date. If the ‘return_all’ argument had been set to FALSE, only new users who were exposed during follow-up (i.e. patients who were flagged 1 in both columns) would be returned.

```r
ex1 <- new_users_fixed(df = synth_presc,
                       drug = "SIMVASTATIN",
                       start_date = "2021-01-01",
                       timeframe = 45,
                       return_all = TRUE,
                       patient_id_col = "patient_id",
                       drug_id_col = "approved_name",
                       presc_date_col = "presc_date")
```

patient_id	new_user	first_presc	exposed
10001	1	2021-01-02	1
10002	0	2021-01-22	1
10003	1	2021-01-29	1
10004	1	2021-01-25	1
10005	NA	NA	0
10006	0	2021-01-08	1

3.7.6 New users – individual start dates (new_users_var)

3.7.6.1 Description

This function classifies patients as new or existing users of the drug(s) of interest at patient-specific start dates. The prescribing and events data are standardised using the ‘tidy_presc’ function, the event dates are joined onto the prescription records by patient ID and records for patients who do not have an event date are removed. A list of the remaining patient IDs is stored for future use. The data are filtered for prescriptions where the drug ID matches the ‘drug’ argument. If the timeframe argument is equal to 0, patients who have a prescription at any time before their listed event date are flagged as existing users. If the ‘timeframe’ argument is not 0, the value is subtracted from the event dates to determine the start of the lookback period for each patient, and patients with prescriptions during that period are flagged as existing users. The patients are then flagged as exposed or unexposed to the drug(s) of interest based on whether or not they have a prescription after their event date, and the date of their first prescription is stored. The flags for existing users are added to the exposure status, and patients who are exposed but are not existing users are marked as new users. If the ‘return_all’ argument is FALSE, the data are then filtered to contain only new users exposed during follow-up. Otherwise, data for all patients is returned, with a flag indicating exposure and new user status and the date of the first prescription for the drug(s) of interest after their event date.

3.7.6.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- df2 – a data frame containing patient event records consisting of at least a patient ID and an event date

- drug – a string containing an identifier for the drug of interest, regular expressions allowed

- timeframe – a number, the desired length of lookback period in days. If set to 0, all prescriptions before the start date will be used; default value is 0

- return_all – logical, if TRUE only records for new users exposed during the period after the start date will be returned, if FALSE records will be returned for all patients with flags indicating if they were exposed and if they are a new user; set to TRUE by default

- patient_id_col, drug_id_col, presc_date_col, ev_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.7.6.3 R code

```
new_users_var <-
  function(df,
           df2,
           drug,
           timeframe = 0,
           return_all = TRUE,
           patient_id_col = "patient_id",
           drug_id_col = "drug_id",
           presc_date_col = "presc_date_x",
           ev_date_col = "ev_date_1",
           date_format) {
    tidy_df <-
      tidy_presc(
        df,
        patient_id_col = patient_id_col,
        drug_id_col = drug_id_col,
        presc_date_col = presc_date_col,
        date_format = date_format
      )
    tidy_ev <-
      tidy_presc(
        df2,
        patient_id_col = patient_id_col,
        ev_date_1_col = ev_date_col,
        date_format = date_format
      )
    tidy_df <- dplyr::left_join(tidy_df, tidy_ev, by = "patient_id")
    tidy_df <- dplyr::filter(tidy_df,!is.na(.data$ev_date_1))
    ids <- tidy_df %>%
      dplyr::select(.data$patient_id) %>%
      dplyr::distinct()
    tidy_df <- tidy_df %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    if (timeframe == 0) {
      uat1 <- tidy_df %>%
        dplyr::group_by(.data$patient_id) %>%
        dplyr::filter(.data$presc_date_x < .data$ev_date_1) %>%
        dplyr::distinct(.data$patient_id) %>%
        dplyr::mutate(new_user = 0)
    } else if (timeframe != 0) {
      uat1 <- tidy_df %>%
        dplyr::group_by(.data$patient_id) %>%
        dplyr::filter(
          .data$presc_date_x < .data$ev_date_1 &
            .data$presc_date_x > .data$ev_date_1 - timeframe
        ) %>%
        dplyr::distinct(.data$patient_id) %>%
        dplyr::mutate(new_user = 0)
    }
    uat2 <- tidy_df %>%
      dplyr::group_by(.data$patient_id) %>%
      dplyr::filter(.data$presc_date_x >= .data$ev_date_1) %>%
      dplyr::summarise(first_presc = min(.data$presc_date_x)) %>%
      dplyr::mutate(exposed = 1)
    uat1 <- dplyr::left_join(ids, uat1, by = "patient_id")
    uat1 <- dplyr::left_join(uat1, uat2, by = "patient_id")
    uat1$new_user[is.na(uat1$new_user) & uat1$exposed == 1] <- 1
    uat1$exposed[is.na(uat1$exposed)] <- 0
    if (return_all == FALSE) {
      uat1 <- uat1 %>%
        dplyr::filter(.data$exposed == 1 & .data$new_user == 1)
      return(uat1)
    } else {
      return(uat1)
    }
  }
```

3.7.6.4 Usage example

In the example below, patients are classified as new users if they did not have a prescription for omeprazole at any point before their first hospitalisation (‘event_1’). The ‘return_all’ argument is set to FALSE, so only data for patients who had a prescription after the event date and none before is returned.

```
ex1 <- new_users_var(df = synth_presc,
                     df2 = synth_events,
                     drug = "OMEPRAZOLE",
                     timeframe = 0,
                     return_all = FALSE,
                     patient_id_col = "patient_id",
                     drug_id_col = "approved_name",
                     presc_date_col = "presc_date",
                     ev_date_col = "event_1")
```

patient_id	new_user	first_presc	exposed
10006	1	2020-12-10	1
10010	1	2020-10-18	1
10015	1	2020-10-04	1
10034	1	2020-12-18	1
10038	1	2020-09-06	1
10051	1	2020-08-17	1

3.7.7 Current vs. past use at event date (uat_recent)

3.7.7.1 Description

This function defines the recentness of patients’ exposure to the drug(s) of interest at an event of interest, based on a user-selected cut-off for current vs. past use. The prescribing and events data are tidied using the ‘tidy_presc’ function, the event dates are appended to the prescribing data based on patient ID and records for patient without an event date are removed. The data are filtered to select only records which match the drug ID provided in the ‘drug’ argument and to remove any prescriptions from after the event date. This also filters out patients who have never had a prescription for the drug(s) of interest. Individual prescriptions are flagged as current use or past use at the event date based on whether or not the prescription date is during the recent use period. The start of the recent use period is calculated by subtracting the value of the ‘timeframe’ argument from the patients’ event date. The number of prescriptions each patient had during the recent and past use periods is calculated, and patients are categorised as current or past users at the event date based on whether or not they have at least 1 recent prescription. The prescription counts and categories are returned as output.

3.7.7.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- df2 – a data frame containing patient event records consisting of at least a patient ID and an event date

- drug – a string containing an identifier for the drug of interest, regular expressions allowed

- timeframe – a number, the desired number of days before the event date where prescriptions should be classified as recent use

- patient_id_col, drug_id_col, presc_date_col, ev_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.7.7.3 R code

```r
uat_recent <- function(df,
                       df2,
                       drug,
                       timeframe,
                       patient_id_col = "patient_id",
                       drug_id_col = "drug_id",
                       presc_date_col = "presc_date_x",
                       ev_date_col = "ev_date_1",
                       date_format) {
  tidy_df <-
    tidy_presc(
      df,
      patient_id_col = patient_id_col,
      drug_id_col = drug_id_col,
      presc_date_col = presc_date_col,
      date_format = date_format
    )
  tidy_ev <-
    tidy_presc(
      df2,
      patient_id_col = patient_id_col,
      ev_date_1_col = ev_date_col,
      date_format = date_format
    )
  uat1 <- dplyr::left_join(tidy_df, tidy_ev, by = "patient_id")
  uat1 <- dplyr::filter(uat1,!is.na(.data$ev_date_1))
  uat1 <- uat1 %>%
    dplyr::filter(grepl(drug, .data$drug_id)) %>%
    dplyr::filter(.data$ev_date_1 >= .data$presc_date_x)
  uat1 <- uat1 %>%
    dplyr::mutate(
      current_flag = dplyr::if_else((
        .data$presc_date_x <= .data$ev_date_1 &
          .data$presc_date_x >= (.data$ev_date_1 - timeframe)
      ),
      1,
      0
      ),
      past_flag = dplyr::if_else(.data$presc_date_x < (.data$ev_date_1 -
timeframe), 1 , 0)
    )
  uat1 <- uat1 %>%
    dplyr::group_by(.data$patient_id) %>%
    dplyr::summarise(
      current_use_flag = max(.data$current_flag),
      n_current = sum(.data$current_flag),
      past_use_flag = max(.data$past_flag),
      n_past = sum(.data$past_flag)
    )
  uat1 <- uat1 %>%
    dplyr::mutate(use_at_event = dplyr::if_else(.data$current_use_flag =
= 1, "current", "past"))
}
```

3.7.7.4 Usage example

In the example below, patients are split into current or past users of atorvastatin at the date of the event of interest (‘event_1’), with the threshold for current use being any prescription in the 90 days before the event. The output provides the number of prescriptions in both the current and past use periods, and a flag indicating which category each user falls into. Patients who did not have any prescriptions for the drug of interest are not included.

```
ex1  <- uat_recent(df = synth_presc,
                   df2 = synth_events,
                   drug = "ATORVASTATIN",
                   timeframe = 90,
                   patient_id_col = "patient_id",
                   drug_id_col = "approved_name",
                   presc_date_col = "presc_date",
                   ev_date_col = "event_1")
```

patient_id	current_user	n_current	past_use_flag	n_past	use_at_event
10016	1	1	0	0	current
10042	1	3	1	1	current
10052	1	2	0	0	current
10061	1	1	0	0	current
10064	1	1	0	0	current
10067	0	0	1	2	past

3.7.8 Two prescriptions within a desired timeframe (uat_gap)

3.7.8.1 Description

This function classifies patients as exposed if they have been prescribed the same drug on two occasions within a defined timeframe. The prescribing records are tidied using the ‘tidy_presc’ function and filtered to remove prescriptions where the drug ID does not match the drug argument. The prescriptions are grouped by patient ID, and the number of days between successive prescriptions is calculated. The records are then checked for instances where the number of days between prescriptions is equal to or less than the ‘timeframe’ argument. If the patient has at least 1 pair of prescriptions matching the exposure threshold they are flagged as exposed, and the dates of the first relevant prescriptions are extracted. Patients who have a prescription for the drug(s) of interest but do not meet the exposure definition are classified as unexposed. If the ‘return_all’ argument is FALSE, the output contains the patient IDs and first and second prescription dates for exposed patients only. If the ‘return_all’ argument is true, data are returned for all patients, including a flag indicating if they were exposed or unexposed, and the prescription dates where applicable.

3.7.8.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- drug - a string containing an identifier for the drug of interest, regular expressions allowed

- timeframe - a number, the maximum length of time in days which should be allowed between successive prescriptions for users to be classified as exposed

- return_all - logical, if TRUE return results for all patients, if FALSE only return results for patients who meet the exposure definition; default value is FALSE

- patient_id_col, drug_id_col, presc_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.7.8.3 R code

```r
uat_gap <- function(df,
                    drug,
                    timeframe,
                    return_all = FALSE,
                    patient_id_col = "patient_id",
                    drug_id_col = "drug_id",
                    presc_date_col = "presc_date_x",
                    date_format) {
  tidy_df <-
    tidy_presc(
      df,
      patient_id_col = patient_id_col,
      drug_id_col = drug_id_col,
      presc_date_col = presc_date_col,
      date_format = date_format
    )
  ids <- tidy_df %>%
    dplyr::select(.data$patient_id) %>%
    dplyr::distinct()
  df1 <- tidy_df %>%
    dplyr::filter(grepl(drug, .data$drug_id))
  df1 <- df1 %>%
    dplyr::group_by(.data$patient_id) %>%
    dplyr::arrange(.data$patient_id, .data$presc_date_x) %>%
    dplyr::mutate(difference = c(0, diff(.data$presc_date_x)))
  df1 <- df1 %>%
    dplyr::mutate(flag = ifelse(dplyr::lead(.data$difference) <= timefra
me, 1, 0))
  df1$flag[is.na(df1$flag)] <- 0
  df1 <- df1 %>%
    dplyr::mutate(exposed = max(.data$flag)) %>%
    dplyr::filter(.data$exposed == 1)
  df2 <- df1 %>%
    dplyr::group_by(.data$patient_id) %>%
    dplyr::filter(.data$flag == 1) %>%
    dplyr::summarise(presc_date_x = min(.data$presc_date_x)) %>%
    dplyr::mutate(presc1 = 1)
  df1 <-
    dplyr::left_join(df1, df2, by = c("patient_id", "presc_date_x"))
  df1$presc1[is.na(df1$presc1)] <- 0
  df1 <- df1 %>%
    dplyr::mutate(presc2 = ifelse(dplyr::lag(.data$presc1) == 1, 1, 0))
  df1$presc2[is.na(df1$presc2)] <- 0
  df2 <- df1 %>%
    dplyr::filter(.data$presc1 == 1 | .data$presc2 == 1) %>%
    dplyr::group_by(.data$patient_id) %>%
    dplyr::summarise(presc1 = min(.data$presc_date_x),
                     presc2 = max(.data$presc_date_x))
  if (return_all == TRUE) {
    df2 <- df2 %>%
      dplyr::mutate(exposed = 1)
    df2 <- dplyr::left_join(ids, df2, by = "patient_id") %>%
      dplyr::select(.data$patient_id,
                    .data$exposed,
                    .data$presc1,
                    .data$presc2)
    df2$exposed[is.na(df2$exposed)] <- 0
  }
  return(df2)
}
```

3.7.8.4 Usage example

In the example below, patients are defined as exposed if they have two prescriptions for simvastatin within 30 days of each other. The output contains records for all patients in the dataset, a flag indicating if they met the threshold for exposure, and the dates of their first and second prescriptions where applicable.

```
ex1 <- uat_gap(df = synth_presc,
               drug = "SIMVASTATIN",
               timeframe = 30,
               return_all = TRUE,
               patient_id_col = "patient_id",
               drug_id_col = "approved_name",
               presc_date_col = "presc_date")
```

patient_id	exposed	presc1	presc2
10001	1	2020-09-29	2020-10-26
10002	1	2020-07-29	2020-08-27
10003	1	2021-01-29	2021-02-24
10004	1	2020-09-22	2020-10-20
10005	0	NA	NA
10006	1	2020-12-09	2021-01-08

3.7.9 Split follow-up into windows (uat_windows)

3.7.9.1 Description

This function splits the period covered by the prescribing data into equal length windows and determines exposure status within each window based on whether or not patients have at least 1 prescription per window. The prescribing and events data are standardised using the ‘tidy_presc’ function, and the follow-up start and end dates are joined to the prescribing data based on patient ID. Patients who do not have two event dates are removed from the dataset, and the data are filtered to remove prescriptions where the drug ID does not match the value of the ‘drug’ argument. If the individual argument is TRUE, the follow-up period is split into periods based on the value of the ‘timeframe’ argument and the patients’ individual start and end dates. If the ‘individual’ argument is FALSE, the minimum start date and maximum start date is used to define the start and end of follow-up for all patients, and this period is divided into windows based on the value of ‘timeframe’. The prescriptions are then split into windows and any prescriptions which fall outside of the follow-up period defined by these dates are removed. The number of prescriptions each patient has within each time window is calculated, and an exposure status is assigned to each time window. If the ‘return_all’ argument is set to TRUE, data are returned for all windows for all patients, indicating the exposure status and number of prescriptions in each window. If the ‘return_all’ argument is FALSE, only windows where patients had at least 1 prescription are returned.

3.7.9.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- df2 – a data frame containing event records consisting of at least a patient ID and the dates of the start and end of follow-up for each patient

- drug – a string containing an identifier for the drug of interest, regular expressions allowed

- individual - logical, if TRUE create individual time windows for each patient based on their start and end dates. If FALSE, use the same windows for all patients based on the first start date and last end date from across all patient records; default value is FALSE

- timeframe - a number, the desired length of each follow-up window in days

- return_all - logical, if TRUE return all windows for all patients, if FALSE only return windows where patients had at least 1 prescription; default value is FALSE

- patient_id_col, drug_id_col, presc_date_col, ev_date_1_col, ev_date_2_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.7.9.3 R code

```
uat_windows <-
  function(df,
           df2,
           drug,
           individual = FALSE,
           timeframe,
           return_all = FALSE,
           patient_id_col = "patient_id",
           drug_id_col = "drug_id",
           presc_date_col = "presc_date_x",
           ev_date_1_col = "ev_date_1",
           ev_date_2_col = "ev_date_2",
           date_format) {
    tidy_df <-
      tidy_presc(
        df,
        patient_id_col = patient_id_col,
        drug_id_col = drug_id_col,
        presc_date_col = presc_date_col,
        date_format = date_format
      )
    tidy_ev <- tidy_presc(
      df2,
      patient_id_col = patient_id_col,
      ev_date_1_col = ev_date_1_col,
      ev_date_2_col = ev_date_2_col,
      date_format = date_format
    )
    uat1 <- dplyr::left_join(tidy_df, tidy_ev, by = "patient_id")
    uat1 <-
      dplyr::filter(uat1,!is.na(.data$ev_date_1) &
                      !is.na(.data$ev_date_2))
    uat1 <- uat1 %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    if (individual == TRUE) {
      uat1 <- uat1 %>%
        dplyr::group_by(.data$patient_id) %>%
        dplyr::mutate(date_group = cut(
          .data$presc_date_x,
          seq(
            min(.data$ev_date_1),
            max(.data$ev_date_2) + timeframe,
            by = timeframe
          )
        )
        %>% as.Date)
    } else if (individual == FALSE) {
      uat1 <- uat1 %>%
        dplyr::mutate(date_group = cut(.data$presc_date_x,
                                       seq(
                                         min(.data$ev_date_1),
                                         max(.data$ev_date_2),
                                         by = timeframe
                                       ))
                      %>% as.Date)
    }
    uat1 <- uat1 %>%
      dplyr::filter(!is.na(.data$date_group))
    uat1 <- uat1 %>%
      dplyr::group_by(.data$patient_id, .data$date_group) %>%
      dplyr::summarise(n_presc = dplyr::n())
    if (return_all == TRUE) {
      uat1 <- uat1 %>%
        dplyr::group_by(.data$patient_id) %>%
        tidyr::complete(date_group = seq.Date(
          from = min(.data$date_group),
          to = max(.data$date_group),
          by = timeframe
        ))
      uat1$n_presc[is.na(uat1$n_presc)] <- 0
      uat1 <- uat1 %>%
        dplyr::mutate(
          end_date = .data$date_group + timeframe - 1,
          exposed = ifelse(.data$n_presc > 0, 1, 0)
        ) %>%
        dplyr::select(
          .data$patient_id,
          .data$date_group,
          .data$end_date,
          .data$exposed,
          .data$n_presc
        )
    } else if (return_all == FALSE){
      uat1 <- uat1 %>%
        dplyr::mutate(
          end_date = .data$date_group + timeframe - 1) %>%
        dplyr::select(
          .data$patient_id,
          .data$date_group,
          .data$end_date,
          .data$n_presc
        )
    }
    return(uat1)
  }
```

3.7.9.4 Usage examples

The examples below highlight two different methods of using this function. In the first example, the ‘individual’ and ‘return_all’ arguments are both FALSE, so the follow-up period is split into the same time windows for each patient, and only time windows where patients had at least 1 prescription are returned. In the second example, both arguments are set to TRUE, so the start and stop dates for follow-up (and therefore the start and end of each window) are individualised for each patient, and all time windows are returned with a flag indicating whether or not the patient was exposed in each.

3.7.9.4.1 Example 1

```
ex1 <- uat_windows(df = synth_presc,
                   df2 = synth_events,
                   drug = "CITALOPRAM",
                   individual = FALSE,
                   timeframe = 90,
                   return_all = FALSE,
                   patient_id_col = "patient_id",
                   drug_id_col = "approved_name",
                   presc_date_col = "presc_date",
                   ev_date_1_col = "start_date",
                   ev_date_2_col = "end_date")
```

patient_id	date_group	end_date	n_presc
10007	2020-07-12	2020-10-09	1
10007	2020-10-10	2021-01-07	1
10009	2020-01-14	2020-04-12	3
10009	2020-04-13	2020-07-11	3
10009	2020-07-12	2020-10-09	3
10009	2020-10-10	2021-01-07	3

3.7.9.4.2 Example 2

```
ex2 <- uat_windows(df = synth_presc,
                   df2 = synth_events,
                   drug = "CITALOPRAM",
                   individual = TRUE,
                   timeframe = 90,
                   return_all = TRUE,
                   patient_id_col = "patient_id",
                   drug_id_col = "approved_name",
                   presc_date_col = "presc_date",
                   ev_date_1_col = "start_date",
                   ev_date_2_col = "end_date")
```

patient_id	date_group	end_date	exposed	n_presc
10007	2020-07-16	2020-10-13	1	1
10007	2020-10-14	2021-01-11	1	1
10009	2020-03-21	2020-06-18	1	6
10009	2020-06-19	2020-09-16	1	3
10009	2020-09-17	2020-12-15	1	2
10009	2020-12-16	2021-03-15	1	3

3.8 Daily-dose related methods

3.8.1 Rationale

The concept of a daily dose - how much of a medication a subject is prescribed for each day - allows for assumptions to be made about the duration or coverage of prescriptions. The DDD is a unit of measurement that is defined by the WHO as ‘the assumed average maintenance dose per day for a drug used in its main indication in adults’. The DDD is a useful unit of measurement for presenting drug utilisation figures, as it allows for assessment of total drug consumption and comparison across patients, patient groups or populations. Depending on the drug in question, the utility of the DDD for assuming prescription duration varies, as it does not always correspond to individual dosing instructions, particularly in cases where drugs are prescribed for multiple indications, where dosage tends to be adjusted over time, or in population groups such as elderly patients or children. The DDD for some drugs may not even be a dosage of a drug which is typically prescribed since it is an average value. The PDD on the other hand, is the actual dose per day that a patient has been instructed to take when they have been prescribed a drug. An assumed coverage period for a prescription can be calculated using the date of the prescription and the number of PDDs dispensed, or in cases where this is not available using assumptions based on the number of DDDs dispensed and a number of DDDs per day. Assumptions on the number of DDDs per day vary across studies, as the concordance between the DDD and days' supply dispensed varies depending on drug class.(209) Daily doses can also be used in a variant of an ever use method, using a minimum number of daily doses dispensed as a threshold for exposure as a categorical variable, or the cumulative number of daily doses dispensed over the follow up period as a continuous exposure variable.

3.8.2 Cumulative daily doses (dd_sum)

3.8.2.1 Description

This function calculates the total number of daily doses of a drug or drugs of interest dispensed per patient by totalling the values of the user-specified daily dose variable in each of the individual prescription record. Records are standardised using the 'tidy_presc' function, and the dataset is filtered on the drug identifier provided ('drug'). Prescriptions are grouped by patient ID, and the total number of prescriptions and the total number of daily doses dispensed are calculated and date of the first prescription for each ID are determined. By default, the function returns this information for each patient who has been prescribed at least 1 daily dose for the drug(s) of interest, but the 'threshold' argument can be adjusted to set a new minimum value.

3.8.2.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier and the number of daily doses dispensed

- drug – a string containing the drug identifier to filter on, regular expressions allowed

- threshold – The function will only return details for patients whose cumulative DDs exceed this value, set to 1 by default

- patient_id_col, drug_id_col, presc_date_col, dd_disp_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.8.2.3 R code

```
dd_sum <- function(df,
                   drug,
                   threshold = 1,
                   patient_id_col = "patient_id",
                   drug_id_col = "drug_id",
                   presc_date_col = "presc_date_x",
                   dd_disp_col = "dd_disp",
                   date_format) {
  tidy_df <-
    tidy_presc(
      df,
      patient_id_col = patient_id_col,
      drug_id_col = drug_id_col,
      presc_date_col = presc_date_col,
      dd_disp_col = dd_disp_col,
      date_format = date_format
    )
  dd1 <- tidy_df %>%
    dplyr::filter(grepl(drug, .data$drug_id))
  dd1 <- dd1 %>%
    dplyr::group_by(.data$patient_id) %>%
    dplyr::summarise(
      n_presc = dplyr::n(),
      total_dds = sum(.data$dd_disp),
      first_presc = min(.data$presc_date_x)
    ) %>%
    dplyr::filter(.data$total_dds >= threshold)
  return(dd1)
}
```

3.8.2.4 Usage example

In the example below, the total number of DDDs of simvastatin dispensed to each patient is calculated. The ‘threshold’ argument is set to 0, so the output contains data for all patients who had at least 1 prescription for simvastatin in the dataset.

```
ex1 <- dd_sum(df = synth_presc,
              drug = "SIMVASTATIN",
              threshold = 0,
              patient_id_col = "patient_id",
              drug_id_col = "approved_name",
              presc_date_col = "presc_date",
              dd_disp_col = "ddd_dispensed")
```

patient_id	n_presc	total_dds	first_presc
10001	12	448.0000	2020-07-05
10002	13	970.6667	2020-05-29
10003	4	298.6667	2020-11-12
10004	8	298.6667	2020-09-22
10006	13	485.3333	2020-05-27
10007	10	746.6667	2020-09-08

3.8.3 Prescription durations (dd_duration)

3.8.3.1 Description

This function calculates a duration (in days) for each prescription. The data are standardised using the 'tidy_presc' function. The database is filtered based on the drug identifier specified, and then durations for each prescription are calculated by multiplying the number of daily doses dispensed by the user-defined factor ('dd_factor'). This value is rounded down to the nearest whole number of days' supply. An end date is then determined for each prescription by adding the duration to the prescribed date - these additional fields are appended to the prescribing data and the modified data are returned as the output.

3.8.3.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier and a number of daily doses dispensed

- drug – a string containing the drug identifier to filter on, regular expressions allowed

- dd_factor - multiplier to be applied to the number of DDs dispensed to calculate the duration and end date for each prescription – adjusting this variable allows for different assumptions regarding the number of DDs per day. Set to 1 by default, and 1 should be used if the DDs dispensed column contains prescribed daily doses

- patient_id_col, drug_id_col, presc_date_col, dd_disp_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.8.3.3 R code

```
dd_duration <- function(df,
                        drug,
                        dd_factor = 1,
                        patient_id_col = "patient_id",
                        drug_id_col = "drug_id",
                        presc_date_col = "presc_date_x",
                        dd_disp_col = "dd_disp",
                        date_format) {
  tidy_df <-
    tidy_presc(
      df,
      patient_id_col = patient_id_col,
      drug_id_col = drug_id_col,
      presc_date_col = presc_date_col,
      dd_disp_col = dd_disp_col,
      date_format = date_format
    )
  dd1 <- tidy_df %>%
    dplyr::filter(grepl(drug, .data$drug_id))
  dd1 <- dd1 %>%
    dplyr::mutate(
      duration = floor(.data$dd_disp * dd_factor),
      end_date = .data$presc_date_x + floor(.data$dd_disp * dd_factor)
    )
}
```

3.8.3.4 Usage example

In the example below, the duration of coverage for individual prescriptions of citalopram is calculated based on a dosage assumption of 0.75 DDDs per day. The output contains a record for each prescription for citalopram from the original data, with the estimate duration and end date appended to the dataset.

```
ex1 <- dd_duration(df = synth_presc,
                   drug = "CITALOPRAM",
                   dd_factor = 0.75,
                   patient_id_col = "patient_id",
                   drug_id_col = "approved_name",
                   presc_date_col = "presc_date",
                   dd_disp_col = "ddd_dispensed")
```

patient_id	presc_date_x	drug_id	dd_disp	duration	end_date
10007	2020-09-18	CITALOPRAM	14	10	2020-09-28
10007	2020-11-15	CITALOPRAM	14	10	2020-11-25
10009	2020-03-21	CITALOPRAM	28	21	2020-04-11
10009	2020-03-21	CITALOPRAM	28	21	2020-04-11
10009	2020-03-23	CITALOPRAM	28	21	2020-04-13
10009	2020-04-17	CITALOPRAM	28	21	2020-05-08

3.8.4 Calculate number of prescribed daily doses dispensed (calculate_pdd)

3.8.4.1 Description

This function calculates the number of prescribed daily doses (PDDs) dispensed based on an individual’s dosing instructions (i.e. how many tablets they should be taking per day). The data are tidied using the ‘tidy_presc’ function, then filtered based on the drug ID provided. The number of prescribed daily doses for each prescription is calculated by dividing the number of tablets dispensed by the instructed number of tablets per day, and this value is appended to the original data as a new column, ‘dd_disp’ and the modified data frame is returned as output.

3.8.4.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a drug identifier the quantity dispensed and the quantity to be taken per day

- drug - a string containing drug identifier to filter for, regular expressions allowed

- drug_id_col, qty_disp_col, qty_per_day_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

3.8.4.3 R code

```
calculate_pdd <- function(df,
                          drug,
                          drug_id_col = "drug_id",
                          qty_disp_col = "qty_disp",
                          qty_per_day_col = "qty_per_day") {
  tidy_df <- tidy_presc(
    df,
    drug_id_col = drug_id_col,
    qty_disp_col = qty_disp_col,
    qty_per_day_col = qty_per_day_col
  )
  dd1 <- tidy_df %>%
    dplyr::filter(grepl(drug, .data$drug_id))
  dd1 <- dd1 %>%
    dplyr::mutate(pdd_disp = .data$qty_disp / .data$qty_per_day)
  return(dd1)
}
```

3.8.4.4 Usage example

The example below contains shows the results of calculating the number of prescribed daily doses of drugs appearing in BNF section 2.12 dispensed based on the dosage instructions provided. The output contains all of the fields of the input data, with the number of PDDs dispensed appended.

```
ex1 <- calculate_pdd(df = synth_presc,
                     drug = "212",
                     drug_id_col = "bnf_section",
                     qty_disp_col = "qty_dispensed",
                     qty_per_day_col = "qty_per_day")
```

patient_id	presc_date	approved_name	drug_id	qty_disp	qty_per_day	pdd_disp
10001	2020-07-05	SIMVASTATIN	212	28	1	28
10001	2020-09-29	SIMVASTATIN	212	28	1	28
10001	2020-10-26	SIMVASTATIN	212	28	1	28
10001	2021-01-02	SIMVASTATIN	212	28	1	28
10001	2021-04-04	SIMVASTATIN	212	28	1	28
10001	2021-06-02	SIMVASTATIN	212	28	1	28

3.9 Persistence

3.9.1 Rationale

Medication persistence can be defined as the period of time from initiation to discontinuation of drug therapy. Discontinuation is the point where a patient can be considered to have stopped taking the drug(s) of interest, based on a lack of evidence of further prescribing or dispensing. At an individual subject level, persistence is typically determined by assessing the number of days from the initial prescription until there a gap in therapy of a specified number of days, with the allowable gap between prescriptions typically depending on clinical relevance and factors related to the number of days' supply provided by an individual prescription - common examples of allowable gaps used in the studies reviewed in Chapter 2 include 14, 30, 60 and 180 days.

The basic refill gap method can be expanded to take account of the estimated coverage of each prescription as determined, for example, by the number of daily doses dispensed. Here, instead of comparing the time between the prescribed date of each prescription the end date of one coverage period is compared with the start date of the next to determine if the allowable gap is exceeded and therefore the patient has discontinued the drug of interest. This can also include extending the coverage period to account for medication stockpiling, where patients have leftover supply of the drug of interest from one prescription when they fill the next prescription before the anticipated end of the previous prescription's supply.

3.9.2 Refill gap only (refill_gap)

3.9.2.1 Description

This function uses a basic refill gap method to determine periods of persistent medication use and points of discontinuation. The prescribing records are standardised using the 'tidy_presc' function and based on the drug identifier entered ('drug'). The prescriptions are grouped by patient ID and drug ID and sorted in ascending date order. The number of days between successive prescriptions is calculated, and a flag is created indicating if this difference is greater than or less than the allowable gap. Each time the gap is greater than the allowable gap, the prescription is tagged as a discontinuation point. Based on these termination points, the function then generates a list of patient IDs, first and last prescription dates in each period of exposure and the length of each exposure (calculated as the difference between the two dates plus the length of the allowable gap). If the 'first_period' argument is set to TRUE, the function removes any periods of use beyond the patient's first exposure to each drug. If the 'threshold' value is set to 0, as by default, all periods are returned. If the value is not equal to zero, only exposures whose length is greater than or equal to the 'threshold' variable are returned - for example, if 'threshold' is 365, only details of subjects' who were persistent for a year or longer after initial prescription will be returned. If the ‘summary’ argument is TRUE, instead of returning individual periods of exposure the function calculates the number of periods of exposure, total number of prescriptions, the dates of the first and last prescriptions and the total length of exposure across all periods for each patient – when the ‘summary’ argument is TRUE, the threshold argument is applied to the total length of exposure, not the length of the individual periods.

3.9.2.2 Arguments

- df - a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- drug – a string containing an identifier for the drug of interest, regular expressions allowed

- gap - the allowable gap between prescriptions for the drug of interest before the patient is considered to have discontinued the medication

- threshold - this value is used to limit the results so that the output contains only exposures with duration greater than or equal to this value, default value is 0

- first_period – logical, if TRUE the function only returns the first period of persistent use for each drug per patient, default value is FALSE

- summary – logical, if TRUE the function returns a summary of all periods of use instead of details of the individual periods; default value is FALSE

- patient_id_col, drug_id_col, presc_date_col - strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format - a string containing the format of the dates used in df

3.9.2.3 R code

```r
refill_gap <-
  function(df,
           drug,
           gap,
           threshold = 0,
           first_period = FALSE,
           summary = FALSE,
           patient_id_col = "patient_id",
           drug_id_col = "drug_id",
           presc_date_col = "presc_date_x",
           date_format) {
    tidy_df <-
      tidy_presc(
        df,
        patient_id_col = patient_id_col,
        drug_id_col = drug_id_col,
        presc_date_col = presc_date_col,
        date_format = date_format
      )
    pers1 <- tidy_df %>%
      dplyr::filter(grepl(drug, .data$drug_id))
    pers1 <- pers1 %>%
      dplyr::group_by(.data$patient_id, .data$drug_id) %>%
      dplyr::arrange(.data$patient_id, .data$presc_date_x) %>%
      dplyr::mutate(difference = c(0, diff(.data$presc_date_x)))
    pers1 <-
      dplyr::mutate(pers1, terminated = ifelse(.data$difference > gap, 1
, 0))
    pers1 <- pers1 %>%
      dplyr::group_by(.data$patient_id, .data$drug_id) %>%
      dplyr::mutate(period = cumsum(.data$terminated)) %>%
      dplyr::select(
        .data$patient_id,
        .data$drug_id,
        .data$presc_date_x,
        .data$difference,
        .data$terminated,
        .data$period
      )
    pers1 <- pers1 %>%
      dplyr::group_by(.data$patient_id, .data$drug_id, .data$period) %>%
      dplyr::summarise(
        first_presc = min(.data$presc_date_x),
        last_presc = max(.data$presc_date_x),
        end_date = max(.data$presc_date_x) + gap,
        n_presc = dplyr::n(),
        duration = (max(.data$presc_date_x) - min(.data$presc_date_x)) +
gap
      )
    if (first_period == TRUE) {
      pers1 <- pers1 %>%
        dplyr::filter(.data$period == 0)
    }
    if (threshold == 0 & summary == FALSE) {
      return(pers1)
    } else if (threshold != 0 & summary == FALSE) {
      pers2 <- pers1 %>%
        dplyr::filter(.data$duration >= threshold)
      return(pers2)
    } else if (threshold == 0 & summary == TRUE) {
      pers2 <- pers1 %>%
        dplyr::group_by(.data$patient_id) %>%
        dplyr::summarise(
          n_periods = dplyr::n(),
          total_n_presc = sum(.data$n_presc),
          first_presc = min(.data$first_presc),
          last_presc = max(.data$last_presc),
          total_length = sum(.data$duration)
        )
    } else if (threshold != 0 & summary == TRUE) {
      pers2 <- pers1 %>%
        dplyr::group_by(.data$patient_id) %>%
        dplyr::summarise(
          n_periods = dplyr::n(),
          total_n_presc = sum(.data$n_presc),
          first_presc = min(.data$first_presc),
          last_presc = max(.data$last_presc),
          total_length = sum(.data$duration)
        ) %>%
        dplyr::filter(.data$total_length >= threshold)
    }
  }
```

3.9.2.4 Usage examples

The examples below highlight the differences in output from the ‘refill_gap’ function based on the use of different arguments. In each example, persistence to simvastatin is determined allowing for a 30-day gap in between prescriptions before discontinuation. In the first example, no additional arguments are used, so the results returned include each individual period of persistent for all exposed patients. In the second example, the ‘threshold’ argument is set to 60, so only periods of use which are at least 60 days long are included in the results. In the third example, the ‘threshold’ argument is set to 60 and the ‘summary’ argument is set to TRUE, so summaries of the overall exposure for each patient are returned for patients whose total duration of persistent use was over 60 days.

3.9.2.4.1 Example 1

```
ex1 <- refill_gap(df = synth_presc,
                  drug = “SIMVASTATIN”,
                  gap = 30,
                  patient_id_col = “patient_id”,
                  drug_id_col = “approved_name”,
                  presc_date_col = “presc_date”)
```

patient_id	period	first_presc	last_presc	end_date	n_presc	duration
10001	0	2020-07-05	2020-07-05	2020-08-04	1	30 days
10001	1	2020-09-29	2020-10-26	2020-11-25	2	57 days
10001	2	2021-01-02	2021-01-02	2021-02-01	1	30 days
10001	3	2021-04-04	2021-04-04	2021-05-04	1	30 days
10001	4	2021-06-02	2021-06-02	2021-07-02	1	30 days
10001	5	2021-07-24	2021-08-23	2021-09-22	2	60 days

3.9.2.4.2 Example 2

```
ex2 <- refill_gap(df = synth_presc,
                  drug = “SIMVASTATIN”,
                  gap = 30,
                  threshold = 60,
                  patient_id_col = “patient_id”,
                  drug_id_col = “approved_name”,
                  presc_date_col = “presc_date”)
```

patient_id	period	first_presc	last_presc	end_date	n_presc	duration
10001	5	2021-07-24	2021-08-23	2021-09-22	2	60 days
10004	2	2021-03-28	2021-05-02	2021-06-01	3	65 days
10006	3	2020-12-09	2021-01-08	2021-02-07	2	60 days
10006	4	2021-02-08	2021-03-26	2021-04-25	3	76 days
10007	0	2020-09-08	2020-10-08	2020-11-07	2	60 days
10007	5	2021-08-16	2021-09-15	2021-10-15	2	60 days

3.9.2.4.3 Example 3

```
ex3 <- refill_gap(df = synth_presc,
                  drug = “SIMVASTATIN”,
                  gap = 30,
                  threshold = 60,
                  summary = TRUE,
                  patient_id_col = “patient_id”,
                  drug_id_col = “approved_name”,
                  presc_date_col = “presc_date”)
```

patient_id	n_periods	total_n_presc	first_presc	last_presc	total_length
10001	9	12	2020-07-05	2022-03-01	355 days
10002	12	13	2020-05-29	2022-02-20	389 days
10003	3	4	2020-11-12	2021-05-22	116 days
10004	5	8	2020-09-22	2021-08-02	213 days
10006	9	13	2020-05-27	2022-01-01	374 days
10007	6	10	2020-09-08	2021-09-15	296 days

3.9.3 Refill gap with coverage (refill_gap_dd)

3.9.3.1 Description

This function determines periods of persistent use of the drug(s) of interest, taking into account the coverage or days' supply given with each prescription and the gap between prescriptions, with the option to consider medication stockpiling. The data are tidied by the 'tidy_presc' function, and then filtered for prescriptions matching the drug identifier provided ('drug'). Then, duration and end date for each prescription are calculated by the 'dd_duration' function, based on the number of daily doses dispensed and the user-selected factor ('dd_disp' and 'dd_factor'). The prescriptions are grouped by patient ID and drug ID, and arranged in date order, and the difference between the end date of one prescription and the start date of the next is calculated sequentially. If the 'stockpile' argument is TRUE, the function checks if the prescribed date for the next prescription is before the predicted end date of the previous prescription and carries over any excess supply. For example, if the new prescription's date is two days before the end date assigned to the previous prescription, the duration of the next prescription is increased by 2. Discontinuation points are determined by checking if the difference between the end of one prescriptions coverage and the date of the next prescription is greater than the allowable gap (gap). Numbered periods of continuous exposure are created based on these discontinuation points. Output is a data frame containing the patient ID, first and last prescription dates, end dates (last prescription date plus duration plus the allowable gap), number of prescriptions and total length of exposure for each period is generated. If the 'first_period' argument is TRUE, the function only returns the first period of exposure for each drug the patient was exposed to. Then, if the value for the 'threshold' argument is zero, this table is returned in full. If another value has been chosen, only periods where the number of days is greater than or equal to the 'threshold' value are returned. If the 'summary' argument is TRUE, instead of returning individual periods of exposure the function calculates the number of periods of exposure, total number of prescriptions, the dates of the first and last prescriptions and the total length of exposure across all periods for each patient – when the 'summary' argument is TRUE, the 'threshold' argument is applied to the total length of exposure, not the length of the individual periods.

3.9.3.2 Arguments

- df – a data frame containing individual prescription records consisting of at least a patient ID, a prescription date, a drug identifier

- drug – a string containing an identifier for the drug of interest, regular expressions allowed

- gap – the allowable gap between prescriptions for the drug of interest before the patient is considered to have discontinued the medication

- threshold – this value is used to limit the results so that the output contains only exposures with duration greater than or equal to this value

- dd_factor – multiplier to be applied to the number of DDs dispensed to calculate the duration and end date for each prescription – adjusting this variable allows for different assumptions regarding the number of DDs per day. Set to 1 by default, and 1 should be used if the DDs dispensed column contains prescribed daily doses

- stockpile – logical, if TRUE any remaining days' supply from one period are carried over to the end of the next period in cases where the second prescription is before the predicted end of the previous supply; default value is FALSE

- first_period – logical, if TRUE the function only returns the first period of persistent use for each drug per patient; default value is FALSE

- summary – logical, if TRUE the function returns a summary of all periods of use instead of details of the individual periods; default value is FALSE

- patient_id_col, drug_id_col, presc_date_col, dd_disp_col- strings, the names of the columns in df containing necessary variables, passed to the tidy_presc function

- date_format – a string containing the format of the dates used in df

3.9.3.3 R code

```
refill_gap_dd <- function(df,
                          drug,
                          gap,
                          dd_factor = 1,
                          threshold = 0,
                          stockpile = FALSE,
                          first_period = FALSE,
                          summary = FALSE,
                          patient_id_col = “patient_id”,
                          drug_id_col = “drug_id”,
                          presc_date_col = “presc_date_x”,
                          dd_disp_col = “dd_disp”,
                          date_format) {
  tidy_df <-
    tidy_presc(
      df,
      patient_id_col = patient_id_col,
      drug_id_col = drug_id_col,
      presc_date_col = presc_date_col,
      dd_disp_col = dd_disp_col,
      date_format = date_format
    )
  pers1 <- tidy_df %>%
    dd_duration(drug = drug, dd_factor = dd_factor)
  if (stockpile == FALSE) {
    pers1 <- pers1 %>%
      dplyr::group_by(.data$patient_id, .data$drug_id) %>%
      dplyr::arrange(.data$patient_id, .data$drug_id, .data$presc_date_x
) %>%
      dplyr::mutate(difference = as.numeric(.data$presc_date_x - dplyr::
lag(.data$end_date)))
    pers1$difference[is.na(pers1$difference)] <- 0
  } else if (stockpile == TRUE) {
    pers1 <- pers1 %>%
      dplyr::group_by(.data$patient_id, .data$drug_id) %>%
      dplyr::arrange(.data$patient_id, .data$drug_id, .data$presc_date_x
) %>%
      dplyr::mutate(stockpile = dplyr::if_else((dplyr::lag(.data$end_dat
e) > .data$presc_date_x),
                                               as.numeric(dplyr::lag(.da
ta$end_date) - .data$presc_date_x),
                                               0
      ))
    pers1$stockpile[is.na(pers1$stockpile)] <- 0
    pers1 <- pers1 %>%
      dplyr::mutate(end_date = .data$presc_date_x + (.data$duration + .d
ata$stockpile)) %>%
      dplyr::mutate(difference = as.numeric(.data$presc_date_x - dplyr::
lag(.data$end_date)))
    pers1$difference[is.na(pers1$difference)] <- 0
  }
  pers1 <-
    dplyr::mutate(pers1, terminated = dplyr::if_else(.data$difference >
gap, 1, 0))
```

```
  pers1 <- pers1 %>%
    dplyr::group_by(.data$patient_id, .data$drug_id) %>%
    dplyr::mutate(period = cumsum(.data$terminated)) %>%
    dplyr::select(
      .data$patient_id,
      .data$presc_date_x,
      .data$drug_id,
      .data$end_date,
      .data$dd_disp,
      .data$difference,
      .data$terminated,
      .data$period
    )
  if(first_period == TRUE){
    pers1 <- pers1 %>%
      dplyr::filter(period == 0)
  }
  pers1 <- pers1 %>%
    dplyr::group_by(.data$patient_id, .data$drug_id, .data$period) %>%
    dplyr::summarise(
      first_presc = min(.data$presc_date_x),
      last_presc = max(.data$presc_date_x),
      end_date = max(.data$end_date) + gap,
      n_presc = dplyr::n(),
      duration = as.numeric(max(.data$end_date) - min(.data$presc_date_x
)) + gap
    )
  if (threshold == 0 & summary == FALSE) {
    return(pers1)
  } else if (threshold != 0 & summary == FALSE) {
    pers2 <- pers1 %>%
      dplyr::filter(.data$duration >= threshold)
    return(pers2)
  } else if (threshold == 0 & summary == TRUE) {
    pers2 <- pers1 %>%
      dplyr::group_by(.data$patient_id, .data$drug_id) %>%
      dplyr::summarise(
        n_periods = dplyr::n(),
        total_n_presc = sum(.data$n_presc),
        first_presc = min(.data$first_presc),
        last_presc = max(.data$last_presc),
        end_of_exposure = max(.data$end_date),
        total_length = sum(.data$duration)
      )
  } else if (threshold != 0 & summary == TRUE) {
    pers2 <- pers1 %>%
      dplyr::group_by(.data$patient_id, .data$drug_id) %>%
      dplyr::summarise(
        n_periods = dplyr::n(),
        total_n_presc = sum(.data$n_presc),
        first_presc = min(.data$first_presc),
        last_presc = max(.data$last_presc),
        end_of_exposure = max(.data$end_date),
        total_length = sum(.data$duration)
      ) %>%
      dplyr::filter(.data$total_length >= threshold)
  }
}
```

3.9.3.4 Usage examples

As in the examples of the use of the ‘refill_gap’ function above, the two examples below define periods of persistent use of simvastatin based on an allowable gap of 30 days between successive prescriptions. In this case, the expected duration of each prescription is calculated using the number of DDDs dispensed and an assumption of 1 DDD per day. In the first example, the ‘stockpile’ argument is set to FALSE, so only the supply of each individual prescription is considered when defining discontinuation. In the second example, ‘stockpile’ is TRUE so any remaining days’ supply from the previous prescription are carried over to the end of the next period. This results in different points of discontinuation for some patients, and therefore different periods of persistent use.

3.9.3.4.1 Example 1

```
ex1 <- refill_gap_dd(df = synth_presc,
                     drug = “SIMVASTATIN”,
                     gap = 30,
                     dd_factor = 1,
                     stockpile = FALSE,
                     patient_id_col = “patient_id”,
                     drug_id_col = “approved_name”,
                     presc_date_col = “presc_date”,
                     dd_disp_col = “ddd_dispensed”)
```

patient_id	period	first_presc	last_presc	end_date	n_presc	duration
10001	0	2020-07-05	2020-07-05	2020-09-10	1	97
10001	1	2020-09-29	2020-10-26	2021-01-01	2	124
10001	2	2021-01-02	2021-01-02	2021-03-10	1	97
10001	3	2021-04-04	2022-03-01	2022-05-07	8	428
10002	0	2020-05-29	2021-01-22	2021-05-06	6	372
10002	1	2021-05-24	2022-02-20	2022-06-04	7	406

3.9.3.4.2 Example 2

```
ex2 <- refill_gap_dd(df = synth_presc,
                     drug = "SIMVASTATIN",
                     gap = 30,
                     dd_factor = 1,
                     stockpile = TRUE,
                     drug_id_col = "approved_name",
                     presc_date_col = "presc_date",
                     dd_disp_col = "ddd_dispensed")
```

patient_id	period	first_presc	last_presc	end_date	n_presc	duration
10001	0	2020-07-05	2020-07-05	2020-09-10	1	97
10001	1	2020-09-29	2021-01-02	2021-03-10	3	192
10001	2	2021-04-04	2022-03-01	2022-05-07	8	428
10002	0	2020-05-29	2022-02-20	2022-06-16	13	778
10003	0	2020-11-12	2021-05-22	2021-09-03	4	325
10004	0	2020-09-22	2020-10-20	2021-01-04	2	134

3.10 Conclusions

This chapter has described the development and contents of the prescribeR package. The primary aim in developing this package was to create a set of functions which could be used by researchers to generate drug exposure variables when preparing routinely collected prescribing data for analysis without having to develop code for each study or project. The structure of these functions makes it relatively simple to test variations of the individual methods or run multiple different methods by changing arguments or using a different function, without having to write and test full scripts. Use of an existing package rather than new code also makes it easier to report on how exposure variables were generated. Instead of having to provide the entire script used or try to summarise the process, reporting can be done by providing the details of the package, function and arguments used, as seen in the usage examples throughout this chapter. Standardising the methods used in this manner then makes it easier for other researchers to reproduce these analyses using other datasets in order to validate the results.

In order to ensure the functions within the prescribeR package were operating as intended, manual validation was performed throughout the development and testing process using both real-world prescribing data and the synthetic datasets included in the package. The synthetic data were used initially to ensure that each function was running as intended, that the output was in the correct format and the results matched what was expected. The benefit of using the synthetic data in this initial step is that the limited number of records makes it easier to manually examine the data and determine the expected results for the full dataset, and quicker to compare these to the test output. After this step is complete and the package appeared to be functioning as intended, the process was repeated on real world data. In this step, the output for several patients was compared against manually derived outputs; in particular, outliers and any seemingly implausible outputs were checked to ensure that they were not produced as a result of errors within the functions.

Although the package was developed using Scottish prescribing data as a reference, the arguments used in the functions are all content neutral. In particular, all of the functions accept regular expressions as the drug argument, meaning that the value entered only has to match the structure of the data stored in the column specified by the user and not a pre-defined standard coding system. Additionally, all of the functions in the prescribeR package use dates in the format specified by the user, not a pre-determined format.

The package is available for download through the repository hosting service GitHub (github.com/amarshall1/prescribeR). In addition to providing a place to store code in a way that allows it to be viewed and used by others, GitHub automatically tracks changes made to the code within the repository and keeps track of different versions, making it easy to revert to a previous version if serious issues arise with new code. GitHub also promotes collaboration, allowing other users to raise issues to highlight bugs in the code and can even create their own branches in the code to add new functionality, which can then be merged into the main branch by the package owner. The package has been assigned a Digital Object Identifier (DOI) (10.5281/zenodo.3834346) which allows others to cite the package if they have used it during their analysis. Further plans for the dissemination of the package include submitting it for inclusion in the CRAN repository, seeking to publish it in a journal and sharing it with colleagues working with routinely collected data within the university.

The next chapter demonstrates the utility of the prescribeR package by using it to prepare exposure variables in a clinical study investigating the rates of drug persistence in a cohort of epilepsy patients.

4 Persistence to anti-epileptic drugs

4.1 Introduction

The previous chapter detailed the contents of the R package prescribeR, written to standardise and simplify the generation of drug exposure variables from routinely collected prescribing datasets. In order to assess the suitability of the package for use in research, testing on a larger, more complex real-world dataset for a defined clinical epidemiology purpose was necessary. Therefore, this chapter discusses the creation of a disease cohort constructed using record-linkage within the NHS Safe Haven environment and its use to demonstrate the utility of the prescribeR package within a clinical research study investigating the rates of persistence to anti-epileptic drugs (AEDs) in a cohort of patients with epilepsy.

4.1.1 Epilepsy

Epilepsy is a common neurological condition typically characterised by recurrent seizures. According to the WHO around 50 million people worldwide have epilepsy and an estimated 5 million people are diagnosed with epilepsy each year.(244) A National Institute for Health and Care Excellence (NICE) guideline on the diagnosis and management of epilepsy, published in 2012, estimated the prevalence of active epilepsy (i.e. patients with continuing seizures or continued need for treatment) in the UK to be 5-10 cases per 1,000 population and estimated the incidence of newly diagnosed epilepsy to be 50 per 100,000 population per year.(245) There is a known association between epilepsy and socioeconomic status, with higher incidence and prevalence amongst less affluent people and in low and middle-income countries.(244)

A number of different epilepsy syndromes exist, typically classified according to clinical features such as type of seizure, neurologic or developmental abnormalities and electroencephalographic (EEG) measurements.(246) Epilepsy syndromes can be split into two broad categories – generalised and partial (or localised) seizure syndromes. In generalised epilepsy syndromes, seizures typically begin simultaneously in both cerebral hemispheres whereas in partial epilepsy syndromes seizures have more localised focal points but can spread to the rest of the brain. There are a number of potential disease mechanisms which can lead to epilepsy, which can be split into structural, genetic, infective, metabolic, immune and unknown causes.

Epilepsy is associated with higher rates of mortality(247) and both physical and mental comorbidities compared to the general population.(248, 249) One study of comorbidities amongst people with epilepsy in Scotland found higher rates of stroke, transient ischaemic attack, chronic liver disease, migraines, learning difficulties and depression amongst people with epilepsy compared to the general population.(248) Another study focusing on the risk of myocardial infarction, stroke and all-cause mortality found that patients with epilepsy exhibited higher risk of all three compared to the general population.(250) Additionally, a Canadian study of psychiatric comorbidity in epileptic patients found there was an increased prevalence of mental health disorders and suicidal ideation in people with epilepsy compared to the general population.(249)

4.1.2 Anti-epileptic drugs

Epilepsy is largely managed through the use of anti-epileptic drug (AED) treatment, with the specific treatment plan typically individualised according to the patients' seizure type, epilepsy syndrome, other medications and comorbidities.(245) When treating epilepsy, the aim is to achieve seizure freedom through the use of AED monotherapy where possible. This often requires atrial-and-error process, and if first-line therapies are ineffective, alternative monotherapies and finally combination therapies are recommended. Approximately two thirds of patients with epilepsy achieve seizure freedom through the use of AEDs.(245)

AEDs act on a diverse set of molecular targets in the central nervous system, with the main aim of modifying the excitability of neurones to block seizure activity.(251) The main mechanisms of action for AEDs are the modulation of voltage-dependent ion channels and the modulation of GABAergic and glutamatergic synaptic activity. In addition to the medications prescribed to prevent seizures from occurring, drugs such as benzodiazepines can be used in emergencies in the case of prolonged or serial convulsive seizures.

The first generation of AEDs (carbamazepine, ethosuximide, phenobarbital, phenytoin, primidone and valproate) were introduced several decades ago, and have complex pharmacokinetic properties and narrow therapeutic windows.(252) Starting in the 1990s, a second generation of AEDs were approved, including gabapentin, lamotrigine, levetiracetam and pregabalin which have generally better safety profiles and wider therapeutic windows(253). Studies investigating the trends in prescribing and utilisation of different AEDs in the UK,(254, 255) Sweden(256), Norway(122) and Australia(257) have shown that while they are still commonly prescribed, the use of older AEDs such as carbamazepine, phenytoin and valproate has decreased over the last 20-25 years, and the use of newer agents such as lamotrigine and levetiracetam as first-line therapies has steadily increased, particularly amongst younger patients.

The NICE guidelines for treatment and management of epilepsy recommend that the decisions regarding which AED to use for each patient should be made based on the patient's seizure type and epilepsy syndrome, as well as their comorbidities and any additional medications they are being prescribed.(245) The large number of AEDs currently available offers a great deal of flexibility to clinicians in specifically tailoring a treatment strategy for each patient, but selecting the best AED can be challenging due to the number of options available. Around 50% of adult patients achieve seizure freedom with their first AED without side effects.(258) The remaining patients will either need to switch to an alternative AED or add a second drug to their regimen, either due to inadequate seizure control or unmanageable adverse effects.

Understanding when and why patients discontinue treatment of different AEDs, and the associations between the duration of persistent use and rates of discontinuation and key demographic and clinical factors could assist in the refinement of prescribing guidance for clinicians on which drugs are more suitable for individual patients with specific characteristics, therefore reducing the need for trial and error in the management of epilepsy and improving patient care.

Although there are a number of published studies investigating rates of adherence to AEDs and the factors which impact adherence, there are few studies focusing on measuring persistence using routinely collected data. One study which investigated persistence to AEDs in a cohort of Taiwanese patients with epilepsy observed a lower risk of non-persistence in patients prescribed oxcarbazepine, valproate, lamotrigine and topiramate compared to patients prescribed carbamazepine, and higher risk of non-persistence in patients prescribed phenytoin compared to those prescribed carbamazepine.(259) The study also found that the mean treatment duration during the first year varied across the drugs investigated, ranging from 218.8 days for gabapentin to 275.9 days for oxcarbazepine.

A second study which examined persistence in a cohort of epilepsy patients in Germany found that less than 50% of cohort patients were persistent to AED treatment after 5 years.(260) Cox proportional hazard regression was used to estimate the risk of discontinuation associated with different demographic and clinical factors. Patients over 60 years of age were at a lower risk of discontinuation than younger patients, and patients with depression were more likely to discontinue therapy than those without depression. Patients prescribed older AEDs (drugs approved before 1980) were found to be more likely to discontinue therapy than patients prescribed newer drugs. Compared to patients who were prescribed valproate, patients prescribed levetiracetam and lamotrigine were at lower risk of non-persistence and patients prescribed gabapentin were at a higher risk. These studies highlight differences in persistence across different drugs, as well as differences across different patient sub-groups, but further work is needed to validate and clarify these findings.

4.1.3 Using routinely collected data for research

As described previously in Chapter 1, everyone who is registered with a GP in Scotland has a CHI number which is associated with all of their encounters with NHS services.(23) The data collected during the provision of these services can be linked using the CHI and anonymised to be used in research through the NHS Safe Havens.

The Safe Havens are a secure environment designed to facilitate the use of anonymised NHS electronic data in research in a way that complies with information governance and maintains patient confidentiality.(261) Currently there are five Safe Havens in Scotland- a national Safe Haven and four regional

Safe Havens located in Aberdeen, Dundee, Edinburgh and Glasgow. They provide access to a range of both nationally collected data and more specialised locally assembled datasets. The approval process for the Greater Glasgow and Clyde (GGC) Safe Haven involves submitting an application containing details of the research question being examined and the datasets required as well as a study protocol, cohort criteria and an analysis plan.(262) Study applications are reviewed by a Local Privacy and Advisory committee consisting of a mixture of clinical staff, academics and members of the public who offer guidance regarding data protection before projects are approved. Researchers are also required to submit documentation regarding their source of funding and current accreditation status. Once the application has been submitted, it is processed and reviewed by the Safe Haven team. If approved, the Safe Haven team then links the required data, and uploads anonymised data to the analysis platform for the researchers to access and analyse.

Data available in the NHS Greater Glasgow and Clyde (NHSGGC) Safe Haven includes birth and death records, inpatient, outpatient, psychiatric, general and maternity admissions and dispensed medicines, as well as more granular datasets such as disease and procedure registers.(262) The ability to connect these data provides an opportunity to investigate issues related to care and outcomes of patients with epilepsy, including use of medications, rates of hospitalisation and the effect of clinical and demographic characteristics on long-term survival. This, in turn, could be used to better understand the relative importance of different factors on disease management and quality of life and, thereby, identify ways to improve the standard of care for patients with epilepsy.

The aim of this chapter is to describe a cohort of patients with epilepsy using linked data covering demographics, hospitalisations, deaths and prescriptions within NHSGGC health board, and to describe the rates of persistence to AEDs in this cohort.

4.2 Methods

4.2.1 Data available in the Safe Haven

The data provided by the Safe Haven were selected based on a list of CHI numbers identified from the NHSGGC's SCI Gateway and TrakCare datasets. SCI Gateway is the system used to manage GP referrals to specialist consultants, and TrakCare is the system used to manage patients' journeys through A&E and outpatient clinics. In total, 10,742 potential patients with epilepsy were identified through referrals to epilepsy outpatient clinics in SCI Gateway or through a record for an epilepsy ICD-10 code in TrakCare (G40.x or G41.x). Before being uploaded to the Safe Haven analysis platform, the individual datasets were linked by CHI number, and then the CHI number was replaced with a project-specific Safe Haven ID number which is matched across the different datasets. CHI numbers are not visible to researchers, as these are confidential patient information.

The available datasets covered demographics, hospitalisations, accident and emergency visits, deaths and dispensed prescriptions. Table 11 summarises the number of records contained within each file before data cleaning and patient selection, the number of patients the data covers (based on the number of unique Safe Haven IDs which appear within each file), the number of records per patient and the date range covered by the data.

Dataset	Number of records	Number of patients	Records per patient (median/IQR)	Date range
Demographics	10,281	10,281	N/A	N/A
Deaths	2,801	2,801	N/A	N/A
Prescribing (PIS)	3,219,093	9,286	259 (389)	01/12/2010 – 31/12/2016 (Prescribed) 30/11/2010 – 31/01/2018 (Dispensed)
Hospitalisations (SMR01)	90,080	9,182	6 (10)	01/12/2010 – 31/12/2016
Accident and Emergency (EDIS)	75,751	9,080	5 (7)	01/12/2010 – 01/04/2017

Table 11 - Summary of the datasets available within the Safe Haven before data cleaning and participant selection

The demography file provided contains demographic characteristics for patients who were residents of the NHS GGC health board taken from the CHI, including date of birth, date of death (if applicable), gender, marital status, postcode sector and Scottish Index of Multiple Deprivation (SIMD) scores. Patients dates of birth were obfuscated by the Safe Haven team prior to the data being uploaded to the analytical platform, as they are considered identifiable patient information – the birth date provided contains the correct month and year, but the day is always set to 15.

The Scottish Morbidity Record (SMR01) contains data on all inpatient and day case admissions to an acute, general hospital. The data include admission and discharge dates, the main condition managed or principal diagnosis, additional conditions or comorbidities, details on the hospital and specialty and significant facility the patient was admitted to, the types of admission (routine or urgent) and discharge (regular discharge, transfer, self-discharge or death), patient management category (day case or inpatient), age on admission, postcode sector and SIMD scores and patients' health board of residence. The conditions or diagnoses associated with each admission were coded according to the International Classification of Diseases (ICD-10).(263)

The Emergency Department Information System (EDIS) contains data on accident and emergency (A&E) attendance taken from the TrakCare system. The data includes admission and discharge dates and times, primary and additional diagnoses or conditions, information on how the patient arrived at A&E and the referral source (e.g. self-referral, GP referral, 999 call), the hospital and A&E department attended, presenting complaint (free text), ICD-10 codes and descriptions, age, sex and ethnicity, cause of injury (free text), information on waiting times within the A&E and vital measurements (e.g. pulse, respiratory rate, blood pressure, temperature).

Death records are obtained from the NRS General Registers Office. The data contains date of death, location of death and ICD-10 codes describing the main and underlying or contributing causes of death.

PIS contains records of prescriptions cashed at local pharmacies within the community in the NHSGGC health board area. The PIS extract for this project included prescription issue and dispensing dates, the approved name, item name and BNF codes for the drug dispensed, the quantity dispensed, the item strength and formulation (tablets, capsules, etc), the number of dispensed items per form, the health board, prescriber type (GP, dentist, nurse, pharmacist etc) and a flag indicating if the patient is a care home resident. Each prescription record contains a BNF code describing exactly what was prescribed. The first section of each BNF item code specifies where the drug appears in the BNF, and the second section specifies the exact product prescribed. An example is provided in Figure 15 below.

Carbamazepine tablets 100mg, generic - 0408010C0AAABAB							
Chapter	Section	Paragraph	Sub-paragraoh	Chemical substance	Product	Strength and Formulation	Generic equivalent
04	08	01	0	C0	AA	AB	AB
Central nervous system	Anti-epileptic drugs	Control of Epilepsy	Control of Epilepsy	Carbamazepine	Generic	100mg tabs	Generic

Figure 15 – Summary of the structure and content of BNF codes assigned to drugs in the pharmacy data using the code for generic carbamazepine 100mg tablets as an example

4.2.2 Data cleaning

4.2.2.1 Demography

Patients not appearing in the demography file are typically patients from other health boards who have received care within NHSGGC. In order to minimise the risk of error due to incomplete data for these patients and to ensure that demographic summaries cover all patients, records for patients who did not appear in this file were removed from the other datasets.

4.2.2.2 Death records

Deaths were recorded in two of the files provided by the Safe Haven: the NRS death records file and the CHI demography file. As these datasets are taken from different sources there are inconsistencies between the two, including deaths which are only recorded in one file or deaths which are recorded in both files but on different dates. In cases where a date of death was recorded in the CHI demography file but not the death records file, a record was added with the cause of death and other information listed as unknown. In cases where patients had a death record but no date recorded in the demography file, or the dates in both files did not match the demography file was adjusted to match the date of death listed in the death records file. After the dates were standardised between the two sources, records were censored at 31/12/2016 to standardise the end of follow-up across all datasets.

4.2.2.3 Hospitalisations and A&E attendances

Duplicate records where all fields were identical were removed. Records where some of the fields were the same but others did not match were retained, as these typically represented additional information related to separate episodes within the same hospital stay, such as movement between different wards. Records were censored at 31/12/2016. Where the discharge date occurred after the recorded date of death, the discharge date was set to the date of death. Flags were added to each record to indicate whether or not an ICD-10 code for epilepsy was listed in the first position or in any position.

SMR01 records represent individual episodes of care - a new record is generated each time a patient is moved between different specialties or significant facilities during a hospital admission. In order to allow for summaries on the number of complete admissions to be generated, an additional dataset was generated containing single records representing continuous inpatient stays. Records were arranged by Safe Haven ID, admission date and discharge date. Records where the admission date was before, or the same as, the previous discharge date were marked as being part of the same stay and records which represented the start of a new stay were flagged. Records were created representing complete stays, containing the admission and discharge dates, the total length of stay, the number of individual episodes that occurred during the stay and flags indicating whether any of the individual records contributing to a continuous stay had an ICD-10 code for epilepsy listed in the first or any position, as above.

4.2.2.4 Prescriptions

Duplicated prescriptions, i.e. records where all fields were identical, were identified and removed from the dataset in order to avoid erroneously increasing the number of recorded prescriptions for these drugs and patients. Prescriptions prescribed or dispensed after 31/12/2016 or after a patients' date of death were removed. A separate file was created containing only the AED prescriptions. Prescriptions for AEDs appearing under alternate BNF codes were adjusted to ensure all AED prescriptions were captured when these records were separated from the other medications.

4.2.3 Cohort selection

The decision was made to include only patients whose diagnosis of epilepsy could be validated based on the data provided. Patients were included in the cohort if they met both of the following criteria:

- be a resident of the NHS GGC health board area as indicated by a presence in the CHI demography file
- At least 1 hospitalisation or A&E attendance with an ICD code for epilepsy in any position
- At least 1 prescription for an AED other than gabapentin

Patients were excluded if they were prescribed gabapentin but had no other AED prescriptions during follow-up as gabapentin monotherapy is typically used to treat neuropathic pain. In order to allow for a standardised lookback period of one year from the start of follow-up for all patients, the start of the study period was defined as 01/01/2012, and patients were only included if they met all of the criteria after this date. Follow-up for patients who met the inclusion criteria started on the date of their first AED prescription or epilepsy-related admission during the follow-up period, whichever occurred first. End of follow-up was defined as either date of death or 31/12/2016, whichever occurred first. After establishing which patients met the inclusion criteria, each of the datasets was filtered to remove data from patients who were excluded. Before analysis, each dataset was then split into two parts covering the follow-up and lookback periods based on individual date of entry into the cohort and end of follow-up dates. If patients had at least 1 AED prescription during the lookback period, they were classified as existing users. The number of different AEDs prescribed to a patient on their index date was used to split patients into those prescribed monotherapy or combination therapy at baseline.

4.2.4 Cohort summary generation

Age at cohort entry was calculated using the date of birth provided in the demography data and the cohort entry date described above. In cases where the obfuscated date of birth resulted in a value of less than 0, age at cohort entry was set as 0. Weighted Charlson comorbidity scores at the start of follow-up were derived from SMR01 records during the 1-year period prior to the start of follow-up using the ‘comorbidity’ package in R, using ICD-10 codes listed in the primary and additional diagnosis fields.(264) The Charlson comorbidity score takes the following comorbidities into account: acute myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic obstructive pulmonary disease, rheumatoid disease, peptic ulcer disease, mild liver disease, diabetes (with or without complications), hemiplegia or paraplegia, renal disease, cancer (any malignancy), moderate or severe liver disease, metastatic solid tumours and HIV/AIDS. When calculating the Charlson comorbidity score, each condition is assigned a score based on the impact it has on a patient’s chance of survival, with additional weighting given to more severe conditions.(265) Diabetes with complications, hemiplegia or paraplegia, renal disease and malignancies are assigned a score of 2, moderate or severe liver disease is assigned a score of 3, metastatic solid tumours and HIV/AIDS are assigned a score of 6 and the remaining comorbidities are assigned a score of 1. The total number of patients in the cohort per calendar quarter was calculated using the number of patients entering the cohort or ending follow-up (either due to death, removal or the end of the follow-up period) during each calendar quarter of the follow-up period. Patients were classified as new AED users at baseline if they had no AED prescriptions in the 1-year period before their index date.

The most common causes of death were determined based on the ICD-10 codes most commonly listed as the primary cause of death. The frequency with which these codes appeared in any position in the cohort death records was also determined. For each patient, the following were calculated:

- the number of continuous inpatient stays, A&E attendances and the total number of admissions

- the number continuous inpatient stays, A&E attendances and total admissions with an ICD-10 code for epilepsy in the first position

- the number continuous inpatient stays, A&E attendances and total admissions with an ICD-10 code for epilepsy in any position

The number of different drug classes (by BNF chapter) and number of unique AEDs (by approved drug name) encashed per patient during follow-up were counted. The most frequently prescribed drug classes (by BNF section) and most frequently prescribed AEDs (by approved name) were calculated, both by the number of prescriptions and the number of patients with at least one prescription during follow-up. The proportion of the total AED prescriptions which were for each of the top 5 AEDs was calculated by quarter across the follow-up period along with the proportion of patients in the cohort who had at least one prescription for these drugs per quarter. To avoid inflating counts for drugs which can be prescribed as two or more forms or doses of the same product at one time, only one prescription per patient per date was counted for each drug class or specific drug. In order to minimise the risk of accidental disclosure of identifying information, any results which apply to fewer than 10 patients were censored as “<10” in line with Safe Haven requirements.

4.2.5 Persistence to AEDs

The number of patients prescribed each AED as their index drug was calculated and split according to whether patients were new or existing users and prescribed monotherapy or combination therapy at baseline. Where patients had more than 1 AED prescribed on their index date, all drugs were counted as index medications and information for those patients was included in more than one group. For patients prescribed more than 1 AED during the study period, prescribing order was identified based on the date of the first prescription for each drug by approved name, and this was used to determine the frequency with which each AED was used as a second or third-line therapy (either as an alternative to their existing therapy or as an add-on therapy).

Persistence to index drugs was determined using a 90-day allowable gap between prescriptions, with patients classified as persistent to a drug if they had not discontinued therapy at 365 days after their index prescription. Persistence was determined for AEDs as a class as well as for individual index drugs. Only patients' first period of persistent use was considered, so if a patient discontinued treatment and was prescribed the drug of interest again after the end of the allowable gap they were still classified as non-persistent.

Persistence variables were generated using the 'refill_gap' function from the prescribeR package, using the arguments displayed below. This command returned the first period of persistent use for each AED for each patient. In order to select for patients' index medication, periods where the start date matched the patient's index date were selected.

```
refill_gap(df = cohort_pharmacy_aed,
           drug = "*",
           gap = 90,
           threshold = 365,
           first_period = TRUE)
```

Rates of add-on and switching during the 365-day period of interest were calculated based on the total number of drugs prescribed at index and at least once over the 365-day follow-up period. Patients were classified as switchers if they were non-persistent to all of their index drugs but had a higher number of different AEDs prescribed overall by the end of the 365-day period compared to baseline. Patients were classified as having add-on therapy if they were persistent to at least 1 index drug and had a higher number of different AEDs prescribed by the end of the 365-day period compared to the number of index medications they were prescribed.

All analyses were carried out in R, using R version 3.5.2 and R Studio version 1.1.463.

4.3 Results

4.3.1 Data cleaning

4.3.1.1 Deaths

After removing patients who did not appear in the demography data, there was a total of 3,094 deaths recorded in either the death records or demography files relating to a total of 10,281 patients. Table 12 below summarises the number of deaths recorded in each file. Of the 2,793 patients who had a death recorded in both files, 12 had two different dates recorded. After standardising the dates and removing 167 deaths which occurred after the censor date, a total of 2,907 deaths remained.

		Death in demography file		
		Yes	No	Total
Death in NRS death records file	Yes	2,793	7	2,800
	No	294	7,187	-
	Total	3,087	-	-

Table 12 - Summary of the dates of death contained in both the demography data and death records

4.3.1.2 Hospitalisations and A&E attendances

The SMR01 dataset contained 639 duplicate records where all fields matched, and no records where the admission date was after 31/12/2016. Records where some fields matched but not all (e.g., if the admission and/or discharge date(s) were the same but different ICD-10 codes were listed or if the admission date was the same, but the discharge dates were different) were retained. Admission dates were compared to the recorded dates of death, and 80 records on 78 patients had an admission or discharge date after the date of death. After selecting for cohort patients and removing records which occurred before or after the follow-up period (described below) a second file was created which combined the individual SMR01 records into 30,093 continuous stays.

The EDIS dataset contained 693 duplicated records where all fields matched and 2,323 records where the admission date was after 31/12/2016, all of which were removed. There were no records where the admission date was after the date of death.

4.3.1.3 Prescriptions

A total of 11,667 (0.4%) prescriptions for patients who were not present in the demography data were removed, along with 56,368 (1.8%) duplicate records, 17,638 (0.5%) records dispensed after the end of the follow-up period and 14,410 (0.4%) prescriptions dispensed after patients' dates of death, leaving 3,119,010 records. A second file containing only AED prescriptions (i.e. drugs with the BNF sub-section code '040801') was created, containing 598,809 records. The BNF codes for 1,734 prescriptions for sodium valproate which were listed in the original data under BNF sub-section '040203' (drugs for mania and hypomania) were adjusted to include these with the other AED prescriptions, bringing the total number of records for AED prescriptions to 592,543.

4.3.1.4 Cohort selection

In total, 5,771 patients met all of the criteria for inclusion in the cohort – Figure 16 below gives a summary of the number of patients who met each of the criteria as they were applied.

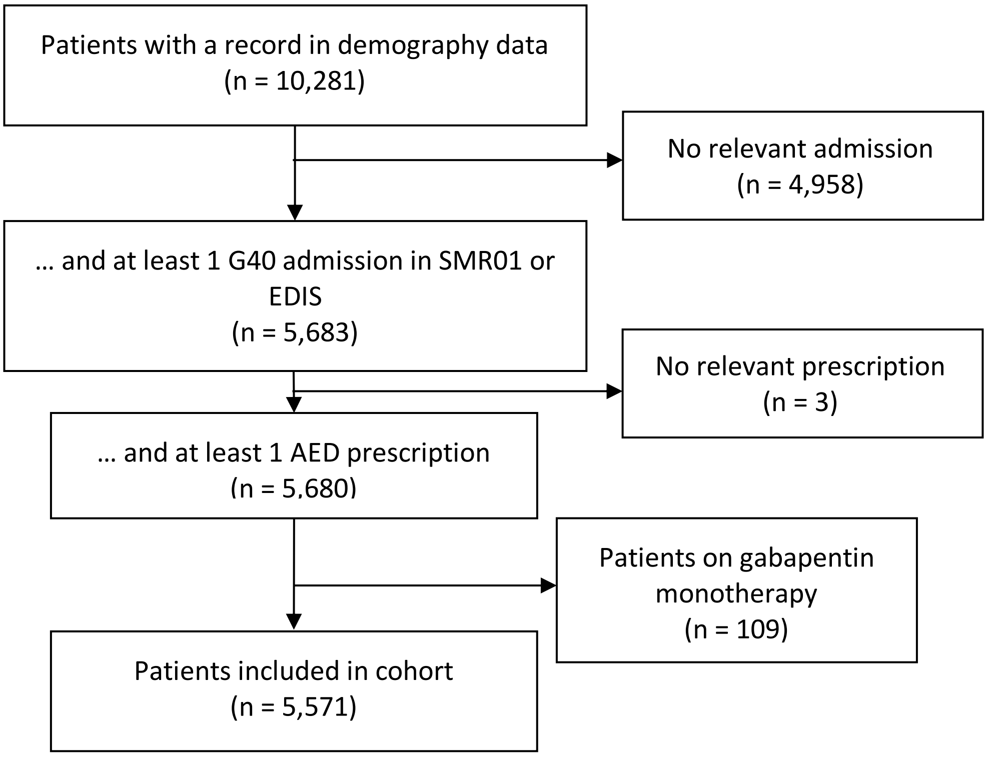

Patients with a record in demography data
(n = 10,281)
No relevant admission
(n = 4,958)
… and at least 1 G40 admission in SMR01 or EDIS
(n = 5,683)
No relevant prescription
(n = 3)
… and at least 1 AED prescription
(n = 5,680)
Patients on gabapentin monotherapy
(n = 109)
Patients included in cohort
(n = 5,571)

Figure 16 - Summary of the number of patients included and excluded at each stage of cohort definition

Once these patients were identified, records for ineligible patients were removed from the datasets, and records from before or after the follow-up period were removed based on individual cohort entry and end of follow-up dates. Table 13 below gives an overview of the contents of each dataset after data cleaning and patient selection – the follow-up files contain all records from that dataset from the patients' cohort entry dates until the end of follow-up, and the lookback files contain the records for the 1 year prior to the patient's cohort entry date if applicable.

File		Number of records	Number of patients	Records per patient (median/IQR)	Date range
Demographics		5,571	5,571	N/A	N/A
Deaths		1,026	1,026	N/A	N/A
PIS – AEDs only	Follow-up	400,073	5,571	52 (73)	04/01/2012 – 30/12/2016
	Lookback	62,129	3859	12 (15)	12/01/2011 – 31/12/2011
PIS – all drugs	Follow-up	1,694,584	5,571	233 (322)	04/01/2012 – 30/12/2016
	Lookback	315,748	5,312	46 (65)	05/01/2011 – 12/12/2016
Hospitalisations (SMR01)	Follow-up	48,451	5,499	5 (8)	01/01/2012 – 30/12/2016
	Lookback	8,687	2,670	2(3)	05/01/2011 – 19/10/2016
Accident & Emergency (EDIS)	Follow-up	38,893	5,377	5(6)	01/01/2012 – 30/12/2016
	Lookback	7,863	3,016	2(2)	08/01/2011 – 23/10/2016

Table 13 - Summary of the data available for cohort patients after data cleaning

Figure 17 shows the total number of patients in the cohort over the duration of the follow-up period, calculated based on the number of patients entering or exiting the cohort during each quarter. The total number of patients being followed-up in a given quarter increased steadily during the first two and a half years of the study period, plateauing around 4,600 patients until declining during the final year of the study period.

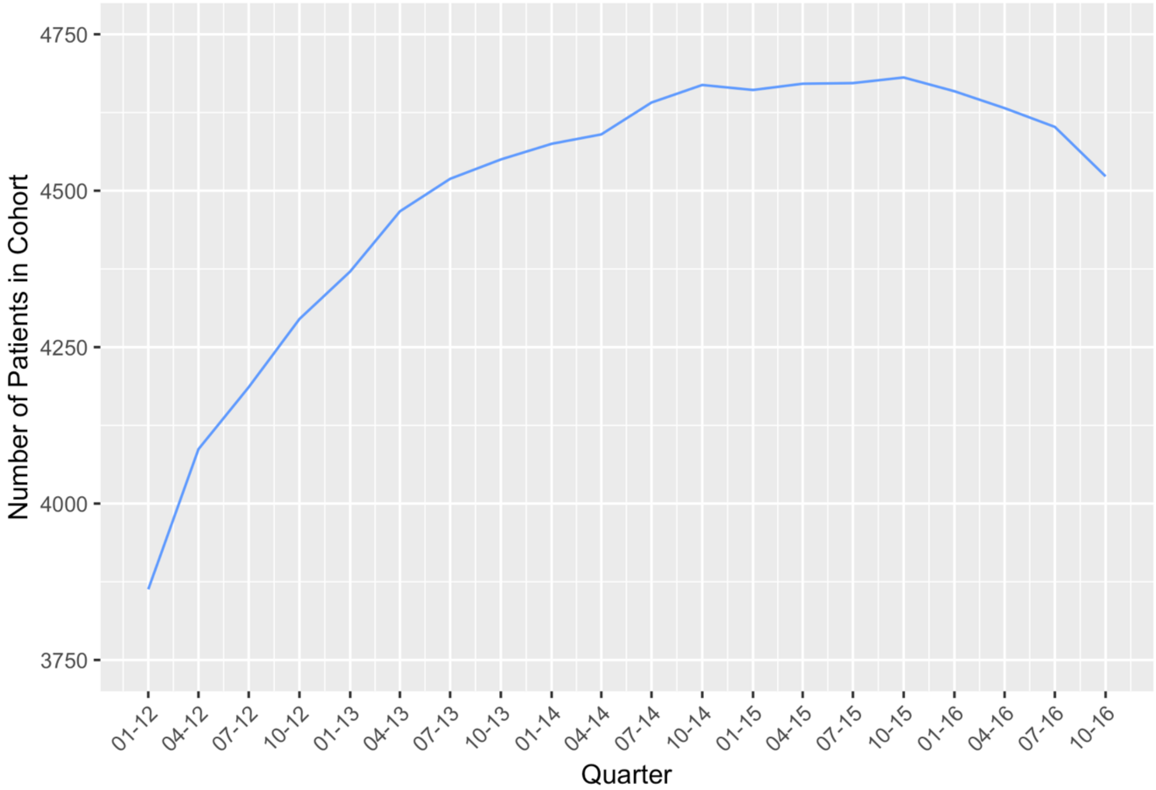


Figure 17 - Summary of the total number of patients in the cohort over the study period by quarter

4.3.2 Demographics

Table 14 summarises the demographic characteristics of cohort patients. The cohort contained a larger number of men than women (54.4% vs. 45.6%), the median baseline age of included patients was 50 years and approximately half (51.7%) of the included patients lived in the most deprived quintile according to the Scottish Index of Multiple Deprivation (SIMD). The median duration of follow-up was 4.83 years, and most patients within the cohort had at least one year of follow-up (93.8%). The majority of patients (82.4%) had a Charlson comorbidity score of 0 at baseline and 18.4% of the included patients died during follow-up.

		All subjects N = 5,571
Gender	Female	2,539 (45.6)
	Male	3,032 (54.4)
Age	0 – 9	327 (5.9)
	10 – 19	348 (6.2)
	20 – 29	500 (9.0)
	30 – 39	627 (11.3)
	40 – 49	921 (16.5)
	50 – 59	989 (17.8)
	60 – 69	837 (15.0)
	70 – 79	652 (11.7)
	80+	370 (6.6)
	Median (IQR)	50 (31)
SIMD Quintile	1 (most deprived)	2,879 (51.7)
	2	956 (17.2)
	3	616 (11.1)
	4	481 (8.6)
	5 (least deprived)	504 (9.0)
	Unknown	135 (2.4)
Charlson comorbidity score (weighted)	0	4,593 (82.4)
	1	508 (9.1)
	2	231 (4.1)
	3+	239 (4.3)
New AED users at baseline	N (%)	1,712 (31.0)
Died	N (%)	1,026 (18.4)
Length of follow-up	Median	4.83 (
	IQR	1.73
	Total	21989.29
	Range	0.03 – 4.99
	Patients with >180d	5,439 (97.6)
	Patients with > 365d	5,228 (93.8)

Table 14 - Summary of the demographic characteristics of cohort patients (n = 5,571)

4.3.3 Outcomes, hospitalisations and deaths

Over the course of follow-up, the median number of hospitalisations per patient was 4, and the median length of hospital stay was 11 days. The median number of ED visits per patient was also 4. The median number of total epilepsy-related admissions per patient was one for epilepsy coded as the main reason for admission and one when coded within the admission for any reason.

Hospitalisations per patient	Median (IQR)	4 (5)
Total LOS per patient (days)	Median (IQR)	11 (35)
A&E attendances per patient	Median (IQR)	4 (6)
Total hospital events per patient	Median (IQR)	8 (10)

Table 15 - Summary of the number of hospital events per patient

		First position	Any position
Epilepsy-related hospitalisations	Median (IQR)	0 (1)	1 (2)
Epilepsy-related A&E attendances	Median (IQR)	0 (1)	0 (1)
Total epilepsy-related hospital events	Median (IQR)	1 (2)	2 (3)

Table 16 - Summary of epilepsy-related hospitalisations and A&E visits for cohort patients, counting admissions where an ICD-10 code for epilepsy was listed in the first position or any position

The most common causes of death amongst cohort patients are listed in Table 17, below. Malignant neoplasm of the bronchus or lung was the most commonly listed primary cause of death, accounting for 5.4%. Epilepsy was the second most common primary cause of death as well as being listed as a contributing cause of death in 17.2% of the death records. The cause of death was only listed as unknown in the 6.6% of death records which were added based on dates of death listed in the demography file during data cleaning.

Most common causes of death (ICD-10 code and description)	N (%) n = 1,026	
	Code listed as primary cause of death	Code listed in any position (primary or contributing)
C439, Malignant neoplasm, bronchus or lung	55 (5.4)	64 (6.2)
G409, Epilepsy, unspecified	45 (4.4)	176 (17.2)
I219, Acute myocardial infarction	43 (4.2)	52 (5.1)
F019, Vascular Dementia	33 (3.2)	61 (5.9)
F03, Unspecified dementia	32 (3.1)	64 (6.2)
I698, Sequelae of other and unspecified cerebrovascular diseases	32 (3.1)	58 (5.7)
R99, Other ill-defined and unspecified causes of mortality	25 (2.4)	26 (2.5)
J690, Pneumonitis due to food and vomit	23 (2.2)	127 (12.4)
J449, Chronic obstructive pulmonary disease, unspecified	22 (2.1)	68 (6.6)
G309, Alzheimer’s disease, unspecified	21 (2.0)	38 (3.7)
I259, Chronic ischaemic heart disease, unspecified	21 (2.0)	109 (10.6)
Unknown	68 (6.6)	N/A
All other codes	606 (59.1)	N/A

Table 17 - Summary of the most common ICD-10 codes listed as the primary or contributing causes of death

4.3.4 Prescriptions

AEDs accounted for 16.7% of the total number of prescriptions encashed by cohort patients, and as it was a requirement for entry into the cohort all patients had at least one prescription for an AED. Analgesics were the second most commonly prescribed drug class both by percentage of prescriptions (7.5% of the total prescriptions) and by percentage of patients (81.4% of patients were prescribed at least one drug from the class).

Antibacterial drugs accounted for a smaller proportion of the total number of prescriptions but were prescribed at least once to over three quarters of cohort patients (79.1%). Drug classes which are used to manage chronic conditions as opposed to acute episodes, including anti-secretory drugs, anti-depressant drugs and lipid-regulating drugs, were amongst the most commonly prescribed classes both by number of patients and number of prescriptions. Topical steroids and emollients were prescribed to large numbers of patients (47.2% and 41.7% respectively) but did not appear on the list of the most commonly prescribed drugs by number of prescriptions, likely due to the quantity prescribed at one time being higher.

Drug class (BNF Section)	Number of prescriptions (%) (N = 1,280,379)	Drug class (BNF Section)	Number of patients (%) (N = 5,571)
Anti-epileptics	213,215 (16.7)	Anti-epileptics	5,571 (100)
Analgesics	95,613 (7.5)	Analgesics	4,535 (81.4)
Anti-secretory drugs	75,593 (5.9)	Antibacterial drugs	4,405 (79.1)
Antidepressant drugs	62,843 (4.9)	Anti-secretory drugs	3,266 (58.6)
Lipid regulating drugs	54,292 (4.2)	Laxatives	2,885 (51.8)
Vitamins	49,185 (3.8)	Topical corticosteroids	2,630 (47.2)
Antiplatelet drugs	47,383 (3.7)	Antidepressant drugs	2,455 (44.1)
Laxatives	42,759 (3.3)	Hypnotics and anxiolytics	2,364 (42.4)
Hypnotics and anxiolytics	39,518 (3.1)	Drugs used in rheumatic disease and gout	2,347 (42.1)
Antibacterial drugs	35,598 (2.8)	Emollient and barrier preparations	2,325 (41.7)

Table 18 - Summary of the most commonly prescribed drugs (grouped by BNF Section) based on total number of prescriptions (counting only 1 prescription per drug class per date for each patient) and number of patients who had at least 1 prescription during follow-up

4.3.4.1 AED prescriptions

Over the course of follow-up, the median number of different AEDs prescribed to cohort patients was 2 per patient and 11.4% of patients were prescribed at least 4 different drugs.

Number of unique AEDs prescribed	N (%) (total = 5,571)
1	2,456 (44.1)
2	1,647 (29.6)
3	829 (14.9)
4	359 (6.4)
5	167 (3.0)
6+	113 (2.0)
Median (IQR)	2 (2)

Table 19 - Summary of the number of different AEDs per patient (by drug name) throughout follow-up

Table 20 summarises the top 10 most commonly prescribed AEDs amongst cohort patients by number of prescriptions and by number of patients with at least one prescription (a complete list of AEDs appearing in the data can be found in Appendix 1). The top three most commonly prescribed AEDs amongst cohort patients were carbamazepine, sodium valproate and levetiracetam, which each accounted for over 15% of the total AED prescriptions and were prescribed at least once to more than 30% of the cohort patients.

Drug	Number of prescriptions (%) (N = 302,961)	Drug	Number of patients (%) (N = 5,571)
Carbamazepine	55,496 (18.3)	Levetiracetam	2,011 (36.1)
Sodium valproate	54,485 (18.0)	Sodium valproate	1,767 (31.7)
Levetiracetam	48,037 (15.9)	Carbamazepine	1,687 (30.3)
Lamotrigine	46,029 (15.2)	Lamotrigine	1,611 (28.9)
Phenytoin	22,480 (7.4)	Gabapentin	680 (12.2)
Gabapentin	12,372 (4.1)	Phenytoin	663 (11.9)
Topiramate	11,073 (3.7)	Pregabalin	550 (9.9)
Pregabalin	10,860 (3.6)	Topiramate	437 (7.8)
Lacosamide	8,474 (2.8)	Lacosamide	415 (7.4)
Phenobarbital	7,390 (2.4)	Clobazam	341 (6.1)

Table 20 - List of the most commonly prescribed AEDs (by drug name) based on total number of prescriptions (counting 1 prescription per drug per day for each patient) and the number of patients who had at least 1 prescription during follow-up

Figure 18 shows the changes in prescribing for the top 5 most commonly prescribed AEDs (carbamazepine, lamotrigine, levetiracetam, phenytoin and sodium valproate) across the study period. At the start of the study period, the most commonly prescribed drug both by proportion of total AED prescriptions and by proportion of patients with at least one prescription was carbamazepine. Over the duration of the study period, lamotrigine and levetiracetam both became more frequently prescribed, with levetiracetam seeing the largest increase and becoming the most frequently prescribed AED during the final year of the study period. The proportion of prescriptions and patients per quarter throughout the study period decreased for both carbamazepine and phenytoin and remained relatively consistent for sodium valproate.

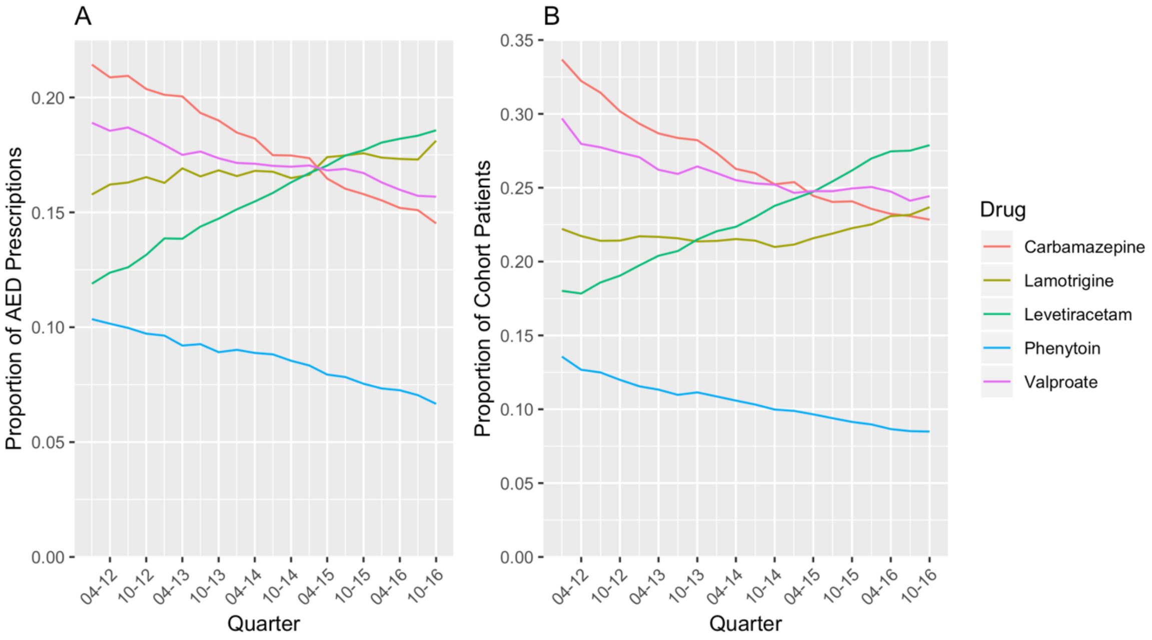
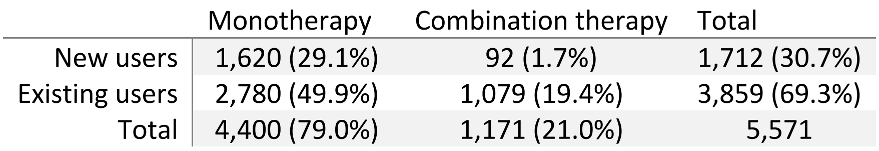


Figure 18 – Summary of changes in AED prescribing during the study period, showing proportion of the total AED prescriptions which were for each of the top 5 most commonly prescribed AEDs (A) and the proportion of cohort patients with at least 1 prescription for each drug (B) per quarter

4.3.5 Persistence to AEDs

A summary of the number of cohort patients who were classified as new and existing users of AEDs and prescribed monotherapy or combination therapy at baseline is shown in Table 21 - Summary of the number of new and existing users of AEDs at baseline and the number of patients initiating monotherapy vs combination therapy. At baseline, more cohort patients were existing users of AEDs than new users based on a 365-day look back period. AED monotherapy was more common than combination therapy, but combination therapy accounted for a larger proportion of existing users vs. new users. Existing users prescribed monotherapy at index made up approximately 50% of the cohort.



	Monotherapy	Combination therapy	Total
New users	1,620 (29.1%)	92 (1.7%)	1,712 (30.7%)
Existing users	2,780 (49.9%)	1,079 (19.4%)	3,859 (69.3%)
Total	4,400 (79.0%)	1,171 (21.0%)	5,571

Table 21 - Summary of the number of new and existing users of AEDs at baseline and the number of patients initiating monotherapy vs combination therapy

4.3.5.1 Index medications

Table 22 contains a summary of the frequency with which each AED was prescribed to cohort patients as their index medication or their second or third medication after the index date. Patients prescribed more than one drug at their index date were included in the totals for multiple drugs. The second or third prescribed drugs either represented alternative or additional drugs prescribed to cohort patients. The cumulative percentages for some drugs are relatively low, as only 26.3% of cohort patients were prescribed 3 or more AEDs during the study period. The AEDs most commonly prescribed to cohort patients as their index therapy were carbamazepine, valproate, levetiracetam, lamotrigine and phenytoin. The most common drugs prescribed as second-line therapies were levetiracetam, gabapentin, lamotrigine, sodium valproate and pregabalin, and the most common third-line therapies were levetiracetam, phenytoin, lamotrigine, lacosamide and gabapentin. Brivaracetam and permapanel were not prescribed to any patients as an index therapy. Stiripentol was only prescribed as the index drug to small number of patients and was not prescribed to any cohort patients as an alternative treatment. Levetiracetam, lacosamide and gabapentin all accounted for higher proportions of second- and third-line therapies than they did index therapy. Carbamazepine accounted for a lower proportion of second- and third line prescribing than index prescribing.

Drug	Index drug		Second drug		Third drug	
			N (%)			
Brivaracetam	0	0%	<10	0.1%	<10	0.1%
Carbamazepine	1,446	26.0%	169	3.0%	47	0.8%
Clobazam	113	2.0%	85	1.5%	71	1.3%
Clonazepam	80	1.4%	68	1.2%	17	0.3%
Eslicarbazepine	<10	0.1%	<10	0.2%	<10	0.1%
Ethosuximide	19	0.3%	<10	0.1%	<10	0.0%
Gabapentin	243	4.4%	310	5.6%	89	1.6%
Lacosamide	90	1.6%	150	2.7%	96	1.7%
Lamotrigine	1,151	20.7%	301	5.4%	97	1.7%
Levetiracetam	1,210	21.7%	605	10.9%	147	2.6%
Oxcarbazepine	48	0.9%	29	0.5%	15	0.3%
Perampanel	0	0%	31	0.6%	68	1.2%
Phenobarbital	170	3.1%	28	0.5%	13	0.2%
Phenytoin	560	10.1%	74	1.3%	113	2.0%
Pregabalin	180	3.2%	196	3.5%	<10	0.0%
Primidone	58	1.0%	14	0.3%	<10	0.1%
Retigabine	<10	0.2%	<10	0.1%	15	0.3%
Rufinamide	<10	0.1%	<10	0.1%	<10	0.0%
Sodium valproate	1,366	24.5%	279	5.0%	84	1.5%
Stiripentol	<10	0.0%	0	0%	0	0%
Tiagabine	<10	0.1%	<10	0.1%	0	0%
Topiramate	225	4.0%	108	1.9%	53	1.0%
Vigabatrin	46	0.8%	<10	0.1%	<10	0.1%
Zonisamide	67	1.2%	30	0.5%	31	0.6%

Table 22 - Summary of the frequencies with which individual AEDs were prescribed as index, second or third drug to cohort patients

Table 23 contains a summary of the number of patients prescribed each AED as their index drug, split according to whether patients were new or existing users and whether they were prescribed monotherapy or combination therapy at baseline. For new users prescribed monotherapy at index, the most commonly prescribed AEDs were levetiracetam, lamotrigine and sodium valproate. For new users prescribed combination therapy, the most commonly prescribed AEDs at index were levetiracetam, carbamazepine and phenytoin. For existing users being prescribed monotherapy at their index date, the most commonly prescribed drugs were carbamazepine, sodium valproate and lamotrigine. For existing users prescribed combination therapy the most commonly prescribed drugs were sodium valproate, carbamazepine and levetiracetam.

Newer AEDs including levetiracetam and lamotrigine were more commonly prescribed as the index medication to new users on monotherapy compared to existing users on monotherapy, and levetiracetam was more commonly prescribed in combination therapy than in monotherapy to both new and existing users at baseline. Older medications including carbamazepine, phenobarbital and phenytoin were more commonly prescribed to existing users on monotherapy at baseline. Phenytoin was more frequently prescribed in combination therapy than monotherapy. Topiramate was more frequently prescribed to existing users on combination therapy than the other groups. Sodium valproate was commonly prescribed across all patient groups but was more common amongst existing users compared to new users and in combination therapy compared to monotherapy.

	New Users				Existing Users			
Drug	Monotherapy		Combination therapy		Monotherapy		Combination therapy	
Carbamazepine	170	10.5%	23	25.0%	824	29.6%	429	39.8%
Clobazam	14	0.9%	<10	<11.0%	34	1.2%	56	5.2%
Clonazepam	13	0.8%	<10	<11.0%	17	0.6%	45	4.2%
Eslicarbazepine	0	0.0%	0	0.0%	<10	<0.4%	<10	<0.9%
Ethosuximide	<10	<0.6%	0	0.0%	<10	<0.4%	<10	<0.9%
Gabapentin	70	4.3%	11	12.0%	63	2.3%	99	9.2%
Lacosamide	<10	<0.6%	<10	<11.0%	13	0.5%	63	5.8%
Lamotrigine	341	21.0%	20	21.7%	440	15.8%	350	32.4%
Levetiracetam	566	34.9%	54	58.7%	232	8.3%	358	33.2%
Oxcarbazepine	<10	<0.6%	0	0.0%	17	0.6%	22	2.0%
Phenobarbital	<10	<0.6%	<10	<11.0%	92	3.3%	69	6.4%
Phenytoin	41	2.5%	24	26.1%	259	9.3%	236	21.9%
Pregabalin	50	3.1%	<10	<11.0%	48	1.7%	73	6.8%
Primidone	<10	<0.6%	<10	<11.0%	21	0.8%	34	3.2%
Retigabine	0	0.0%	0	0.0%	<10	<0.4%	<10	<0.9%
Rufinamide	<10	<0.6%	0	0.0%	0	0.0%	<10	<0.9%
Sodium valproate	271	16.7%	23	25.0%	633	22.8%	439	40.7%
Stiripentol	0	0.0%	0	0.0%	0	0.0%	<10	<0.9%
Tiagabine	0	0.0%	<10	<11.0%	<10	<0.4%	<10	<0.9%
Topiramate	26	1.6%	<10	<11.0%	60	2.2%	135	12.5%
Vigabatrin	20	1.2%	<10	<11.0%	<10	<0.4%	15	1.4%
Zonisamide	<10	<0.6%	<10	<11.0%	<10	<0.4%	50	4.6%

Table 23 - Summary of AEDs prescribed as index drugs to cohort patients, split by new and existing users and whether patients were prescribed monotherapy or combination therapy at baseline, number of patients and percentage of each patient sub-group reported

4.3.5.2 Persistence to AEDs

A summary of the overall rates of persistence to AEDs at 365 days is shown in Table 24. The rate of persistence to AEDs as a class was 5.6% higher than persistence to any individual index drug. This difference represents patients who discontinued their index treatment but were still considered persistent when all AEDs were considered together and were therefore still receiving some form of AED treatment at the end of the 365-day period of interest. Approximately 25% of cohort patients had a change in their AED therapy over the 365-day period, having either switched to a different AED or having had another AED added to their treatment during the 365-day period following their index prescription.

	N (%) (N = 5,571)
Patients persistent to AEDs as a class	3,572 (64.1%)
Patients persistent to any index drug	3,277 (58.5%)
Patients persistent to all index drugs	3,102 (55.7%)
Patients with an add-on therapy	698 (12.5%)
Patients who switched therapy	666 (12.0%)

Table 24 - Summary of the rates of persistence to AEDs as a class and any or individual AEDs at 365-days after index prescription, and the rates of add-on prescribing or switching within the first 365 days

A summary of overall persistence to index AEDs split based on whether patients were new or existing users and prescribed monotherapy or combination therapy at baseline is shown in Table 25. For patients prescribed combination therapy and all patients, the rate of persistence to any index AED and the rate of persistence to all index AEDs are reported. Persistence rates were generally higher amongst existing users than new users and patients on combination therapy rather than patients prescribed monotherapy. Persistence was highest among patients who were existing users and started follow-up on combination therapy, although rate of persistence to all index drugs was lower than to any individual drug, as 15% of these patients discontinued at least 1 index medication.

	Monotherapy	Combination therapy		Total	
		Any drug	All drugs	Any drug	All drugs
New users	661 (40.8%)	48 (52.2%)	35 (39.1%)	709 (41.4%)	696 (40.7%)
Existing users	1707 (61.4%)	861 (79.8%)	699 (64.8%)	2568 (66.5%)	2406 (62.3%)
Total	2368 (53.8%)	909 (77.6%)	734 (62.7%)	3277 (58.8%)	3102 (55.7%)

Table 25 – Summary of overall persistence rates for any or all index drugs at 365 days after index prescription, with the number of patients persistent and the percentage of each group persistent to any index therapy and all index therapies reported

The rates of persistence to index AEDs broken down by individual drug are summarised in Table 26. Persistence rates ranged from 33.3% - 100% amongst all drugs. The drugs which were prescribed least, including rufinamide, tiagabine esclicarbazepine, ethosuximide and retigabine were outliers with either particularly low or high rates of persistence. Amongst the more commonly prescribed (N>10) drugs, persistence ranged from 50.6% (gabapentin) to 71.2% (phenobarbital), with a median persistence rate of 61.4%.

Index drug	Patients persistent N (%)	
Carbamazepine	933	64.5%
Clobazam	79	69.9%
Clonazepam	56	70.0%
Eslicarbazepine	<10	37.5%
Ethosuximide	<10	47.4%
Gabapentin	123	50.6%
Lacosamide	56	62.2%
Lamotrigine	696	60.5%
Levetiracetam	662	54.7%
Oxcarbazepine	33	68.8%
Phenobarbital	121	71.2%
Phenytoin	368	65.7%
Pregabalin	92	51.1%
Primidone	40	69.0%
Retigabine	<10	33.3%
Rufinamide	<10	100.0%
Sodium valproate	821	60.1%
Tiagabine	<10	100.0%
Topiramate	131	58.2%
Vigabatrin	24	52.2%
Zonisamide	36	53.7%

Table 26 –Summary of the rates of persistence to individual AEDs at 365 days after index prescription

The rates of persistence to index AEDs by drug split by patient sub-group are summarised in Table 27. Persistence rates were generally higher for existing users than for new users, and also for individual drugs when prescribed as part of combination therapy as opposed to monotherapy.

Drug	New users				Existing users			
	Monotherapy		Combination therapy		Monotherapy		Combination therapy	
	N	%	N	%	N	%	N	%
Carbamazepine	60	35.30%	11	47.80%	527	64.0%	335	78.1%
Clobazam	<10	35.70%	<10	44.40%	24	70.6%	46	82.1%
Clonazepam	<10	46.20%	<10	80.00%	11	64.7%	35	77.8%
Eslicarbazepine	-	-	-	-	<10	50.0%	<10	25.0%
Ethosuximide	<10	25.00%	-	-	<10	33.3%	<10	75.0%
Gabapentin	15	21.40%	<10	54.50%	28	44.4%	74	74.7%
Lacosamide	<10	42.90%	<10	100.00%	<10	46.2%	40	63.5%
Lamotrigine	161	47.20%	<10	45.00%	276	62.7%	250	71.4%
Levetiracetam	236	41.70%	22	40.70%	137	59.1%	267	74.6%
Oxcarbazepine	<10	55.60%	-	-	11	64.7%	17	77.3%
Phenobarbital	<10	50.00%	<10	80.00%	59	64.1%	56	81.2%
Phenytoin	14	34.10%	10	41.70%	174	67.2%	170	72.0%
Pregabalin	13	26.00%	<10	44.40%	29	60.4%	46	63.0%
Primidone	<10	0.00%	<10	50.00%	13	61.9%	26	76.5%
Retigabine	-	-	-	-	<10	0.0%	<10	50.0%
Rufinamide	<10	100.00%	-	-	-	-	<10	100.0%
Sodium valproate	126	46.50%	<10	30.40%	368	58.1%	320	72.9%
Stiripentol	-	-	-	-	-	-	<10	0.0%
Tiagabine	-	-	<10	100.00%	<10	100.0%	<10	100.0%
Topiramate	<10	19.20%	<10	50.00%	32	53.3%	92	68.1%
Vigabatrin	<10	30.00%	<10	33.30%	<10	62.5%	12	80.0%
Zonisamide	<10	12.50%	<10	100.00%	-	-	31	62.0%

Table 27 – Summary of rates of persistence to individual AEDs at 365 days after index prescription split by new and existing users and whether patients were prescribed monotherapy or combination therapy at baseline

4.4 Discussion

The aim of this chapter was to describe a cohort of patients with epilepsy and to examine the rates of persistence to AEDs. The results showed the rates of prescribing different AEDs as index therapies or additional therapies later in treatment, as well as highlighting variation in the rates of prescribing different AEDs and in 1-year persistence to AED therapy by drug. The results also demonstrate differences in both frequency of prescribing of different AEDs and rates of 1-year persistence to AEDs between new and existing users and patients prescribed the drugs as amonotherapy or in combination with other AEDs. This chapter also demonstrated the utility of the prescribeR package in preparing data for clinical research.

A total of 5,770 patients with epilepsy were identified using epilepsy-related prescriptions and hospital records representing 21,989.29 person-years of follow-up, with a median follow-up length of 4.83 years per patient. Over the course of the follow-up period 18.4% of the cohort patients died. This is a relatively high percentage of patients, but given the number of patients over 65 in the cohort and the lower life expectancy for patients with epilepsy this is not entirely unexpected. The level of comorbidity amongst cohort patients was assessed using the Charlson comorbidity score, calculated using records of hospitalisations in the year prior to cohort entry. Most patients (82.4%) scored 0, suggesting a low prevalence of serious comorbid conditions amongst cohort patients, a finding which is not in line with other research on patients with epilepsy.(248) This may be due to a number of factors, including the limited length of the lookback period meaning older hospitalisations are not accounted for, the use of only hospitalisation records to ascertain comorbidity and the specific subset of conditions included in the Charlson comorbidity score. Further comorbidities could likely be identified through the use of primary care data or other indices of comorbidity, but these data are not currently available for this cohort. Another potential method of expanding the range of comorbid conditions considered would be to use pharmacy data as a proxy to identify other conditions including conditions such as diabetes, asthma, depression and anxiety which often do not require hospitalisation and for which there are drugs specific to the condition. When using pharmacy records to define comorbidity, the use of medications must be specific enough in order to minimise risk of misclassification, particularly in cases where drugs have multiple indications.

Prescribing of multiple AEDs per patient was common in this cohort - over the course of the study period, over half of the included patients were prescribed two or more different AEDs, and 10% were prescribed 4 or more. The most commonly prescribed AEDs and trends in the utilisation of the top AEDs observed for this cohort are in line with those observed in the literature. (122, 254-257) The top 5 AEDs (carbamazepine, valproate, levetiracetam, lamotrigine and phenytoin) accounted for almost three quarters (74.8%) of all AED prescriptions across the follow-up period, and around 30% of the study patients had at least one prescription for levetiracetam, valproate, carbamazepine or lamotrigine. Overall, carbamazepine was the most commonly prescribed AED across the follow-up period by number of prescriptions, but the trends over time showed a decrease in the use of older drugs (carbamazepine and phenytoin) while the number of prescriptions and number of patients prescribed newer drugs increased, particularly levetiracetam which became the most commonly prescribed AED by both number of prescriptions and number of patients during the last year of the study. This may be due to these drugs having fewer side effects, and is likely in line with these newer drugs becoming available in generic forms; levetiracetam became available as a generic medication in 2011(266), followed by lamotrigine's chewable and dispersible formulations in 2012.(267) As shown in Chapter 4, the rates of prescribing of levetiracetam and lamotrigine generally increased across the follow-up period for this cohort.

There was variation in the rates of persistence across AEDs; across all cohort patients, 58.8% were persistent to at least 1 of their index medications, with persistence across individual drugs ranging from 50.6% to 71.2% for drugs prescribed to more than 10 patients at index. When comparing the different patient sub-groups investigated in this chapter, new users had lower rates of persistence than existing users, particularly new users on monotherapy. Existing users on combination therapy had the highest rates of persistence to at least 1 index therapy, as well as to most of the individual AEDs. There are a number of factors which could explain these differences. Existing users may be more used to managing their condition and continuing to take their medications as prescribed, have greater disease severity and therefore a greater reliance on AEDs to control seizures, or be on medications which are already shown to be effective in controlling their seizures. New users who are non-persistent to their index medications, on the other hand, may be switching to other drugs on physician's advice. Combination therapy users may be more likely to persist due to increased disease severity; combination therapy is a last resort option for physicians, and the recommendation is to try alternative monotherapies before considering adding a second medication to a patient's regime.

The use of the prescribeR package made the generation of the large quantity of persistence variables summarised in this chapter relatively simple, as the majority of the work was done through one function call. This allowed for comparison of the number of patients prescribed different drugs and the rates of persistence to these drugs across a whole class with minimal additional coding. In this case, additional code was required to filter the drug use periods returned by the function to exclude non-index medications for each patient, a feature which could be a useful addition to the next version of the persistence functions in the package.

With the required data cleaning and processing complete, the epilepsy cohort described in this chapter can be used to perform analyses for a range of different research questions related to prescribing and outcomes for patients with epilepsy. CHI linkage also offers potential to extend the cohort data by linking to additional datasets available within the Safe Haven platform including outpatient clinic attendance (SMR00), mental health admissions (SMR04) or laboratory test data. The established process for cleaning the existing datasets also allows for extension of the cohort to include additional patients or to extend the study follow-up period for the existing data as more data are captured and made available. For example, updating the analysis to include the most recently collected data would allow for assessment of the impact of changes made in 2018 to the guidance on the prescribing of sodium valproate to female patients of childbearing age.(268)

The established cohort is not without limitations. The definition of epilepsy used in the inclusion criteria relies on the patient having been hospitalised or having attended A&E and the hospital event having been correctly recorded as epilepsy related. This means that there may be patients who have less severe cases of epilepsy or focal epilepsy syndromes where patients do not experience motor symptoms who are not included in the cohort as a result of not being hospitalised. The more severe cases are still of interest when investigating issues around management of epilepsy through AED prescribing and the challenges of managing patients with multimorbidity. Finally, studies investigating issues around drug utilisation typically distinguish between new and existing users of the drug(s) of interest. Using the current cohort definition and lookback period, only 31.0% of included patients were considered new users of AEDs at baseline (i.e. they had no prescriptions for AEDs in the one year prior to cohort entry). Initially, the intention was to use patients whose diagnosis of epilepsy had previously been validated through the use of GP records and epilepsy-related referrals to neurology outpatient clinics. However, we were unable to verify that the data provided in the Safe Haven were for these patients exclusively, so the decision was made to include the requirement for an epilepsy-related hospital event to verify that the included patients were epileptic.

Although there were a large number of cohort patients overall, there were relatively small numbers of patients prescribed certain drugs as their index medications, including ethosuximide, esclicarbazepine, retigabine, rufinamide and tiagabine, all of which were prescribed to less than 10 cohort patients at baseline. This means that the persistence rates reported for these drugs may not be a reliable estimate of persistence in a larger population, but as this is a population-based cohort these low prescribing rates indicate these drugs are infrequently used in the community.

Additionally, while a 1-year lookback period was used to identify patients who were existing users of AEDs at baseline, there is potential that some patients are misclassified and are in fact existing users who have been previously non-persistent. This is particularly true of new users who were prescribed combination therapy at baseline, as combination therapy is only usually used in cases where a number of different monotherapies have been tried and have failed to adequately manage the patient’s seizures. Additionally, for patients who were classified as existing users of AEDs at baseline, no adjustments were made for whether or not they had been taking their index drug(s) during the lookback period, or the length of time they had already been persistent if they had. For patients on combination therapy, the drugs were only investigated individually, so persistence to specific combinations of AEDs was not considered. In cases where patients were non-persistent to one drug but were found to have switched to a second medication during the 365-day period of interest, persistence to the second drug was not measured. This means there was no differentiation between patients who had discontinued AED therapy completely and patients who had persisted on a second drug beyond the initial measurement of persistence to AEDs as a class. Finally, although the differences in persistence across different patient groups were assessed, this chapter did not consider the impact of factors such as disease severity, comorbidity, demographic factors on comorbidity or the variable potential for adverse events across different drugs.

The results in this chapter have demonstrated the potential utility of the prescribeR package as a tool for generating drug exposure variables in clinical research using large, routinely collected prescribing databases, and have demonstrated variation in the levels of persistence to different AEDs in a cohort of patients with epilepsy. In the next chapter, this cohort will be used to test the prescribeR package and assess the effect that varying the definition of drug exposure has on the observed associations in an exemplar clinical research question.

5 Comparison of exposure quantification methods

5.1 Introduction

The systematic review of pharmacoepidemiological research detailed in Chapter 2 highlighted a range of methods for quantifying drug exposure based on routinely collected data. These methods varied in the level of detail they provided, and even within the different classes of methods identified there was variation in the exact definition of exposure used between studies. Each of the assumptions required when quantifying exposure has the potential to introduce exposure misclassification into the study which would limit the validity and accuracy of the results obtained. There is evidence showing that the drug of interest,(209, 269) and the structure, source and potential weaknesses or limitations of the data being used(26, 34, 198, 227, 233) should be taken into account when deciding how exposure should be quantified.

Although the systematic review highlighted some of the relative merits and limitations of these methods, there were a limited number of studies assessing the impact of altering the definition of drug exposure on the results observed in pharmacoepidemiological research. The aim of this chapter is to demonstrate the effects of varying the definition of exposure on the estimated number of patients exposed and the estimated duration of treatment episodes for a range of different AEDs and other drug classes, and to demonstrate how these changes impact the observed association between exposure to levetiracetam and all-cause mortality in the previously defined cohort of epilepsy patients.

5.2 Methods

5.2.1 Study population

As described in Chapter 4, in order to be eligible for Inclusion in the cohort, patients had to be a resident of the NHSGGC health board area, as indicated by a presence in the CHI demography file, have at least one epilepsy-related hospital admission recorded in SMR01 or A&E attendance recorded in the EDIS data, and have at least one prescription for an AED other than gabapentin. The study period started on 01/01/2012 and ended on 31/12/2016, and patients began follow-up on the date of their first epilepsy-related admission, A&E attendance or AED prescription during the study period, whichever occurred first. All data used were collected within the NHSGGC health board area.

5.2.2 Effect of exposure definition on the estimated number of exposed patients

The exposure definitions assessed in this chapter were chosen based on the common methods and thresholds identified in Chapter 2.

5.2.2.1 Ever use

Patients were defined as being exposed if they had at least the threshold number of prescriptions for the drug(s) of interest to assess the differences between the number of patients exposed as the threshold was increased from 1 to 10 prescriptions. The following drugs/drug groups were assessed:

- AEDs as a class (i.e. prescriptions for any drug with the BNF sub-section code ‘040801’)
- The top 5 most commonly prescribed AEDs (by approved drug name); carbamazepine, sodium valproate, lamotrigine, levetiracetam, phenytoin
- Other commonly prescribed drug groups (by BNF section code) – analgesics (0407), lipid-regulating drugs (0212), antibacterial drugs (0501) and anti-depressants (0403)

The structure of the R function calls used to determine which patients meet each definition is shown below – the ‘drug’ and ‘threshold’ values were adjusted to provide all of the desired exposure variables.

```r
ever_use(df = cohort_pharmacy_aed,
        drug = "*",
        threshold = 1)
```

Some of the drugs of interest are often prescribed in multiple forms at one time - for example, if patients are on a higher dose, they may be prescribed two tablets of different strengths to make up the prescribed dose. When comparing patients on different doses and prescribing across different drugs of interest, not accounting for these records could introduce bias by inflating some patients' prescription counts, or the prescription counts for certain drugs. In the example shown in Figure 19, counting both the 10mg and 20mg prescriptions the third patient has received would increase their total prescription count, therefore misclassifying their exposure level.

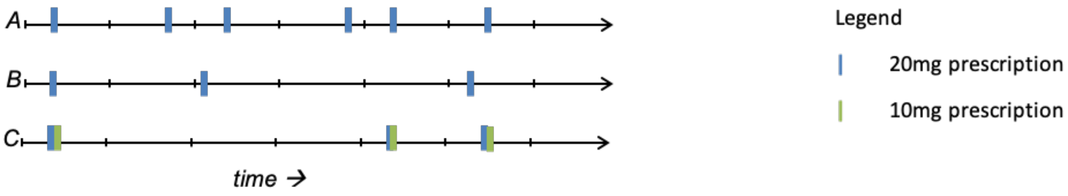


Figure 19 - Example showing how drugs prescribed in multiple doses can affect prescription counts

In order to assess the impact of combining these records on the number of patients exposed at each threshold, the number of patients exposed when all prescriptions were taken into account was compared to the number of patients exposed when only counting one prescription per date. The number of patients who had at least one prescription for any drug was used as the baseline to calculate the proportion of patients who met the other thresholds to allow for clearer comparisons between drugs or drug classes. The structure of the function call used to derive these variables is shown below – the 'drug' and 'threshold' were adjusted to match each exposure definition.

```
ever_use(df = cohort_pharmacy_aed,
        drug = "*",
        threshold = 1,
        flatten = TRUE)
```

5.2.2.2 Use at time

Three methods of defining exposure based on use at a specific time point, or within a specific time frame, were investigated. The impact of using different lookback periods on the number of patients who were defined as new users of each drug of interest prescribed AEDs at baseline. To be classified as a new user at baseline, a patient needed to have at least one prescription for the drug during follow-up and no prescriptions for the drug of interest within the period of interest before their index date. The proportion of patients ever prescribed the drug during follow-up who were classified as new users using each lookback duration was calculated and compared across drugs. The function below was used to determine patients' exposure status, with the 'drug' and 'timeframe' variables were adjusted to match each exposure definition.

```
new_users_var(df = cohort_pharmacy_aed,
              df2 = cohort_follow_up,
              drug = "*",
              timeframe = 7)
```

Next, the number of patients who were classified as current users of each drug of interest at death, was calculated using different allowable timeframes before their date of death - the proportion of patients who died and had ever had a prescription who met the exposure definition was calculated.

```
uat_fixed_events(df = cohort_pharmacy_aed,
                 df2 = cohort_deaths,
                 drug = "*",
                 timeframe = 7,
                 forward = FALSE)
```

Finally, the number of patients who met the exposure definition of two prescriptions for the drug of interest within a desired number of days of one another was determined using different allowable gaps, and the proportions of the number of patients who had at least 2 prescriptions on different dates and met each exposure definition were calculated. The 'drug' and 'timeframe' variables function call below were adjusted to derive the necessary exposure variables

```
uat_gap(df = cohort_pharmacy_aed,
        drug = "*",
        timeframe = 7)
```

For each of these three methods, the allowable timeframes tested were 7, 14, 30, 60, 90, 180, 270 and 365 days. The number of patients meeting the definition of exposure for AEDs as a class and each of the top 5 most commonly prescribed AEDs was assessed.

5.2.2.3 Persistence

The effect of varying the duration of the allowable gap between prescriptions on rates of persistence at 180 days and 365 days after initiating treatment and the length of time to first discontinuation was assessed. The number of patients who were considered persistent to the drug of interest 180 and 365 days from their first prescription was determined. The rate of persistence at both time points was assessed for AEDs as a class and each of the top 5 most commonly prescribed AEDs. When considering AEDs as a class, the patient was considered persistent even if they switched to a different AED or started on an additional AED therapy.

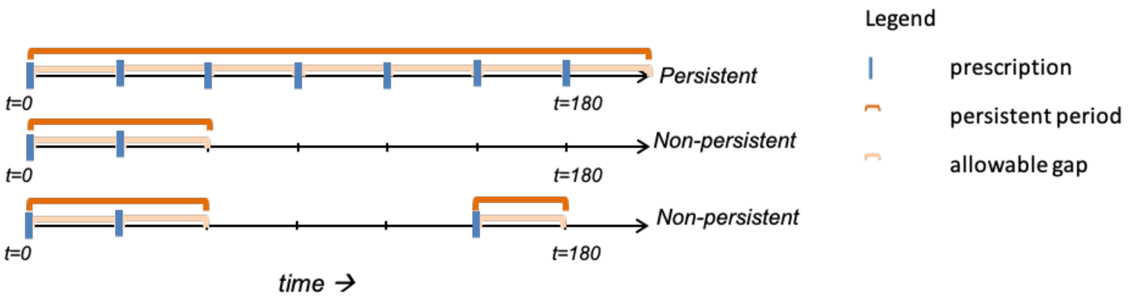


Figure 20 – Example application of the definition of 180-day persistence applied to cohort patients

The allowable gap between prescriptions was varied, with 30, 45, 90, 120- and 180-day allowable gaps tested. The proportion of patients who had at least 1 prescription and had not discontinued treatment at 180 and 365 days was calculated. Once patients were identified as non-persistent, any further use within the 180- and 365-day period but after the allowable gap was ignored – an example of how 180-day persistence was assessed is shown in Figure 20. The length of time persistent until the first point of discontinuation using each allowable gap length was determined and summarised across the cohort. Figure 21 shows an example of how time to discontinuation was calculated for cohort patients passed on individual prescriptions.

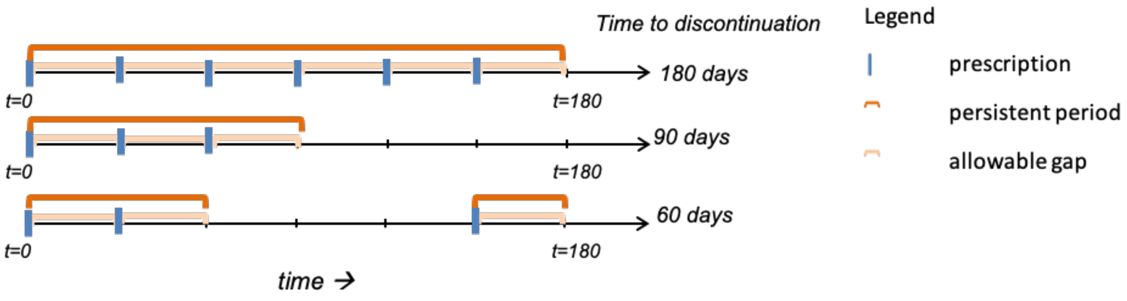


Figure 21 - Example of how time to discontinuation was calculated for cohort patients

The function call below was used to identify patients' first period of persistent use and the 'drug', 'gap' and 'threshold' arguments were adjusted to provide the measurements described above.

```
refill_gap(df = cohort_pharmacy_aed,
           drug = "*",
           gap = 30,
           first_period = TRUE)
```

5.2.3 Effect of exposure definition on observed exposure-outcome associations

The association between levetiracetam exposure and all-cause mortality was calculated using a number of different exposure definitions as part of a series of Cox proportional hazards models. Cohort patients were classified as exposed or unexposed to levetiracetam based on a number of different time-fixed and time-varying exposure definitions. The index date for all patients was their cohort entry date.

The baseline characteristics (age, sex, SIMD quintile, length of follow-up and Charlson comorbidity scores) of the exposed and unexposed patients were compared to determine if any substantial differences existed between the two groups at baseline. For this comparison, patients were considered exposed to levetiracetam if they had at least 1 prescription during follow-up, as this included all patients who would appear in the exposed group in any of the definitions used for the models described below.

In the time-fixed definitions, exposure was defined as a binary variable which did not change during the follow-up period. The definitions used were;

- 1 or more prescription (i.e. ever versus never use),
- 2 or more prescriptions,
- 2 prescriptions within 30 days,
- 2 prescriptions within 90 days,
- 2 prescriptions within 180 days.

The necessary exposure statuses for each patient were defined using previously described calls to the ‘ever_use’ and ‘uat_gap’ functions. Follow-up for exposed patients began on the date of their first levetiracetam prescription in the first model and the date of their second levetiracetam prescription for the remaining models.

In the time-varying definitions, prescription records were assessed across the follow-up period to define exposure status at multiple points. The simplest method used here was the legacy effect, whereby patients were considered unexposed until their first prescription for levetiracetam, and then exposed until the end of follow-up. Additional methods involved splitting the follow-up period into windows of equal length and defining exposure based on whether or not patients had at least one prescription within each window, and defining persistent time using different allowable gap lengths. The definitions used were;

- Legacy effect
- Follow-up split into 30, 90- or 180-day windows
- Persistent time defined with a 30, 90 and 180-day allowable gap

The exposure status and first prescription date for each patient for the legacy effect model were taken from a previously described call to the ever_use function. The following function calls were used to define the second and third group of exposure periods, with the ‘timeframe’ and ‘gap’ variables adjusted respectively. Follow-up for all patients began on their index date.

```
uat_windows(df = cohort_pharmacy_aed,
            df2 = cohort_follow_up,
            drug = "LEVETIRACETAM",
            individual = TRUE,
            timeframe = 30)
```

```
refill_gap(df = cohort_pharmacy_aed,
           drug = "LEVETIRACITAM",
           gap = 30,
           first_period = FALSE)
```

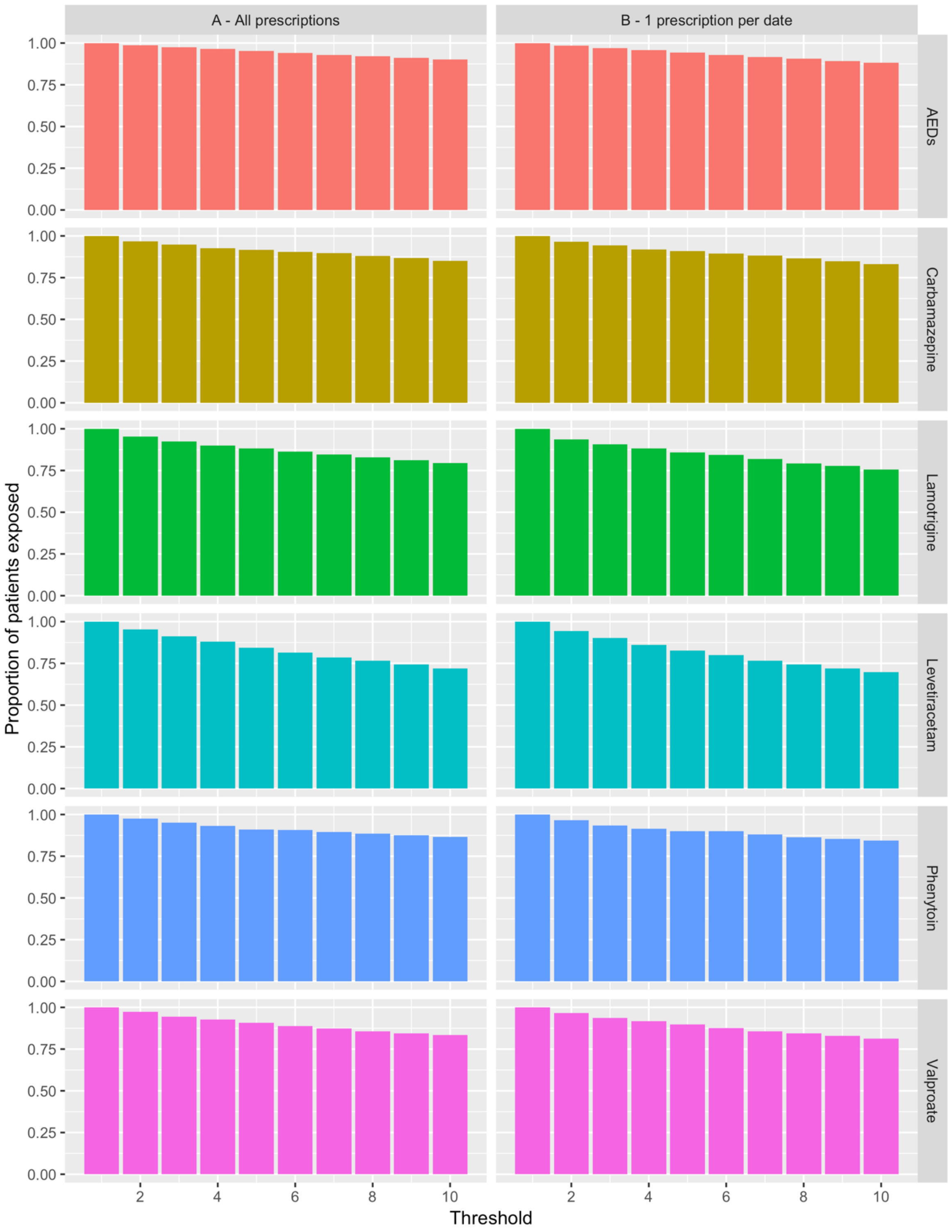
In each model, the reference group consisted of patients who did not meet the threshold for exposure to levetiracetam - these patients all had at least 1 prescription for another AED, as this was required for entry into the cohort. In addition to levetiracetam exposure, all of the models were adjusted for age, sex and comorbidity score at baseline. Cox proportional hazard regression models were used for all exposure methods to calculate hazard ratios and 95% confidence intervals (CI). The aim was not to establish an association between levetiracetam use and all-cause mortality but to highlight any systematic differences in observed association which occur as a result of varying exposure definition, so other limitations in the use of these models were not explored. As the same cohort and variables are used in each model, they should have the same degree of additional confounding.

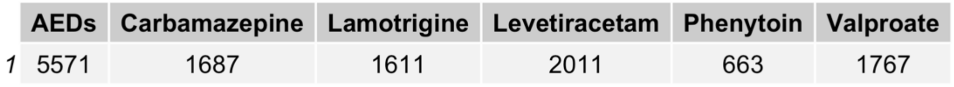
All analyses were carried out in R, using R version 3.5.2 and R Studio version 1.1.463. Exposure variables were generated using the prescribeR package.

5.3 Results

5.3.1 Ever use

A summary of the changes in the proportion of patients who are considered exposed to the drugs of interest as the threshold for exposure increases can be seen in Figure 22, along with the total number of patients with at least 1 prescription for AEDs as a class or each individual drug. For all of the drugs of interest, at least 70% of patients were still considered exposed with a threshold of 10 prescriptions using both methods. When grouping all AEDs as a class and counting all individual prescriptions, 90.1% of the patients with 1 prescription are still considered exposed at the 10-prescription threshold. This value decreased to 88.1% when only considering 1 prescription per patient per day. Amongst the individual drugs, phenytoin retained the largest number of exposed patients as the exposure threshold increased (86.6%/84.3% at 10 prescriptions, depending on method), whereas levetiracetam experienced the largest decrease in exposed patients (79.5%/75.6%). For all drugs the number of patients classified as exposed was lowered by switching from including all prescriptions to just one per patient per day. Using the 10-prescription threshold, the difference between the proportion of patients defined as exposed by the two methods ranged from 1.9% to 3.8%.

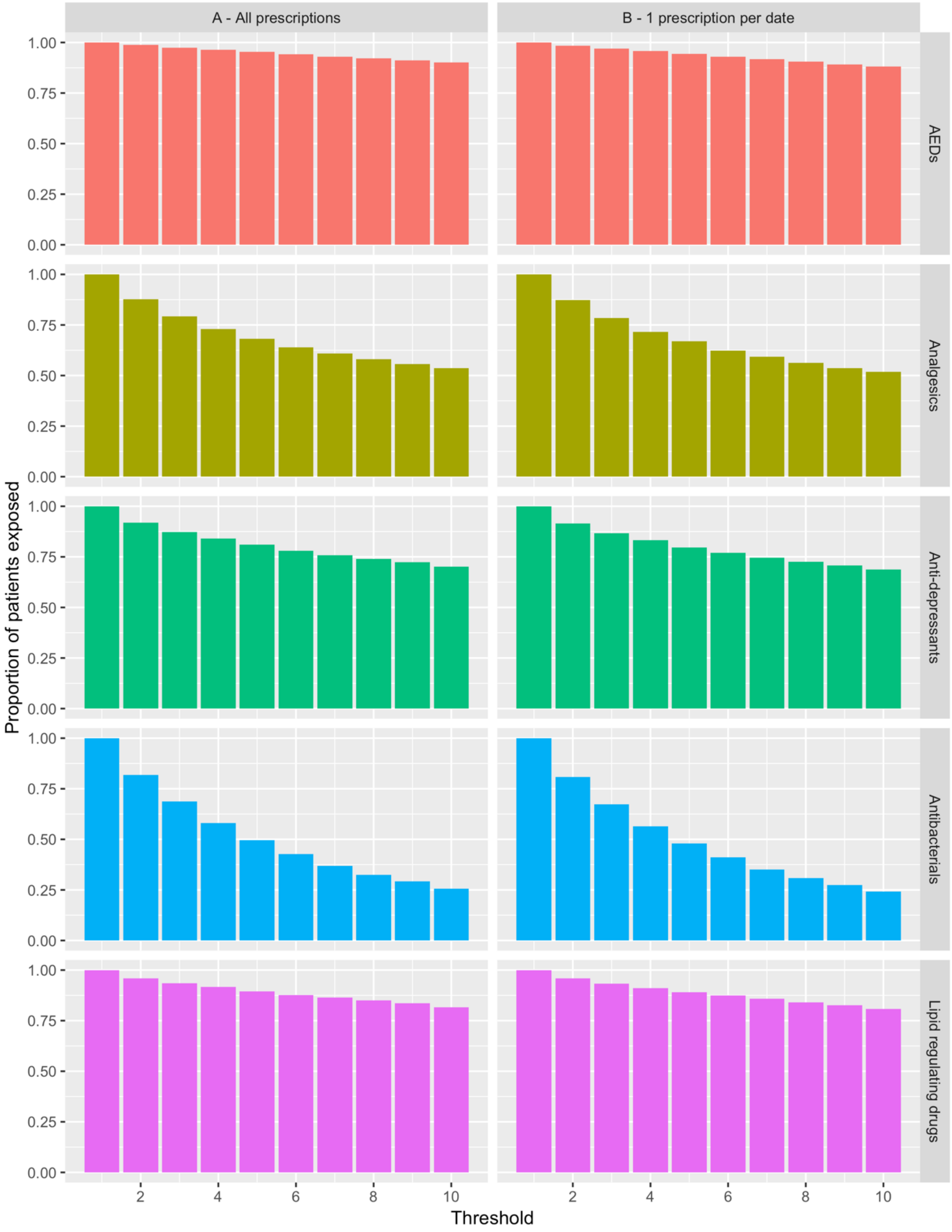




	AEDs	Carbamazepine	Lamotrigine	Levetiracetam	Phenytoin	Valproate
1	5571	1687	1611	2011	663	1767

Figure 22 - Proportion of patients with 1 prescription for AEDs as a class and each of the top 5 AEDs who meet the threshold for exposure when including all prescriptions (A) or 1 prescription per date (B), with the total number of patients who had at least 1 prescription for each shown in the table

Figure 23 shows a summary of the changes in the proportion of patients exposed to different classes of drugs as the threshold number of prescriptions is increased, as well as the total number of patients with at least 1 prescription for each drug class. AEDs have the highest retention rate of patients with increasing threshold. There is larger variation across the different drug classes than across individual drugs within the same class as described above. A smaller reduction in the exposed patient group is seen for drugs used to treat chronic conditions or for long-term prevention (AEDs, lipid-regulating, anti-depressants). The largest reduction in the size of the exposed group was seen for antibacterials - using an exposure threshold of 10 prescriptions, only 24.2% of the patients who had at least 1 prescription would still be considered exposed. As with the individual drugs there is a difference between the two methods of counting prescriptions, but the largest difference between two points for any of the drug classes was 2.0% for AEDs.

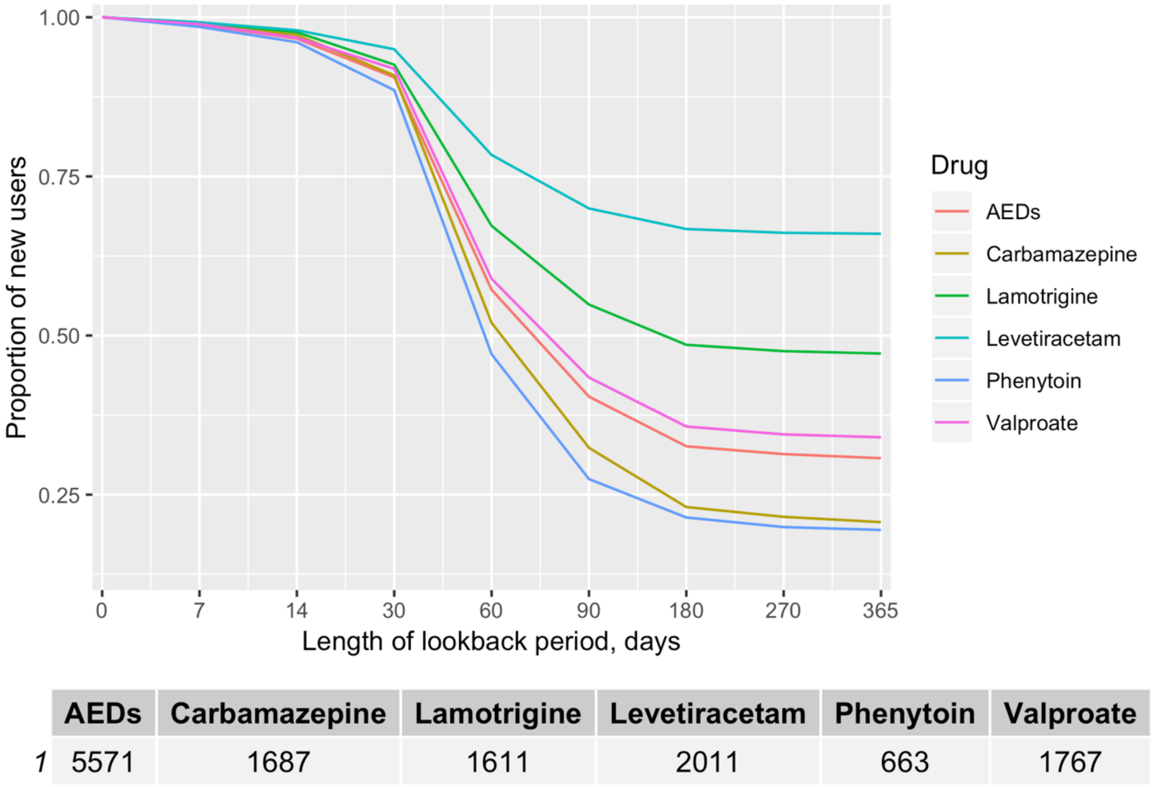


	AEDs	Anti-depressants	Analgesics	Antibacterials	Lipid-regulating drugs
1	5571	2455	4353	4405	1881

Figure 23 - Proportion of patients with 1 prescription for different classes of who meet the threshold for exposure when including all prescriptions (A) or 1 prescription per date (B), with the total number of patients who had at least 1 prescription for each shown in the table

5.3.2 Use at time

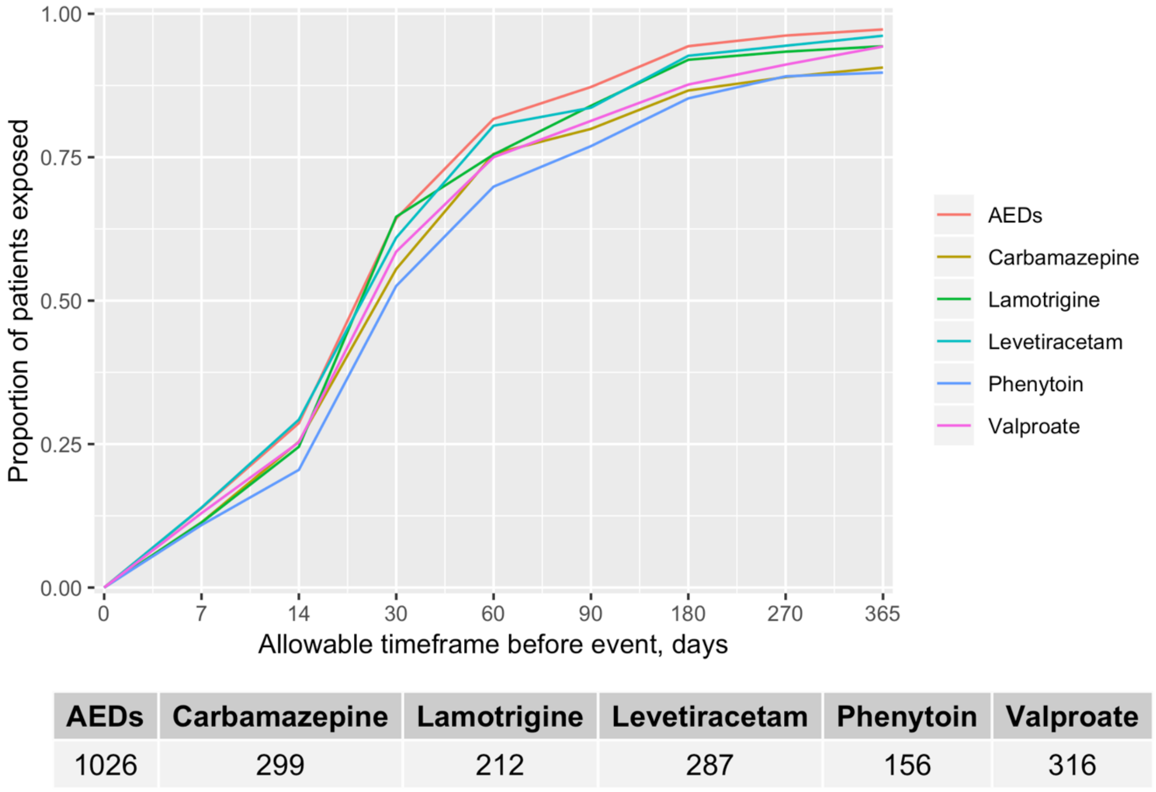
Figure 24 summarises changes in the proportion of patients ever exposed to AEDs as a class and the most commonly prescribed AEDs, who were classified as new users based on varying periods of lookback from their index date. There was variation across the drugs of interest. When using a 30-day lookback period, 89% of phenytoin users and over 90% of the patients exposed to the other drugs during follow-up were considered new users. Using a 365-day lookback period there is more variation in the proportion of new users – 66% of levetiracetam users were still classified as new users, only 19% of phenytoin users were new users and 31% of patients were new users when considering AEDs as a class as opposed to individual drugs.



	AEDs	Carbamazepine	Lamotrigine	Levetiracetam	Phenytoin	Valproate
1	5571	1687	1611	2011	663	1767

Figure 24 - Proportion of patients who have at least 1 prescription for drug(s) of interest who are classified as new users at cohort entry based on varying lengths of lookback period, with the total number of patients who had at least 1 prescription during follow-up shown in the table

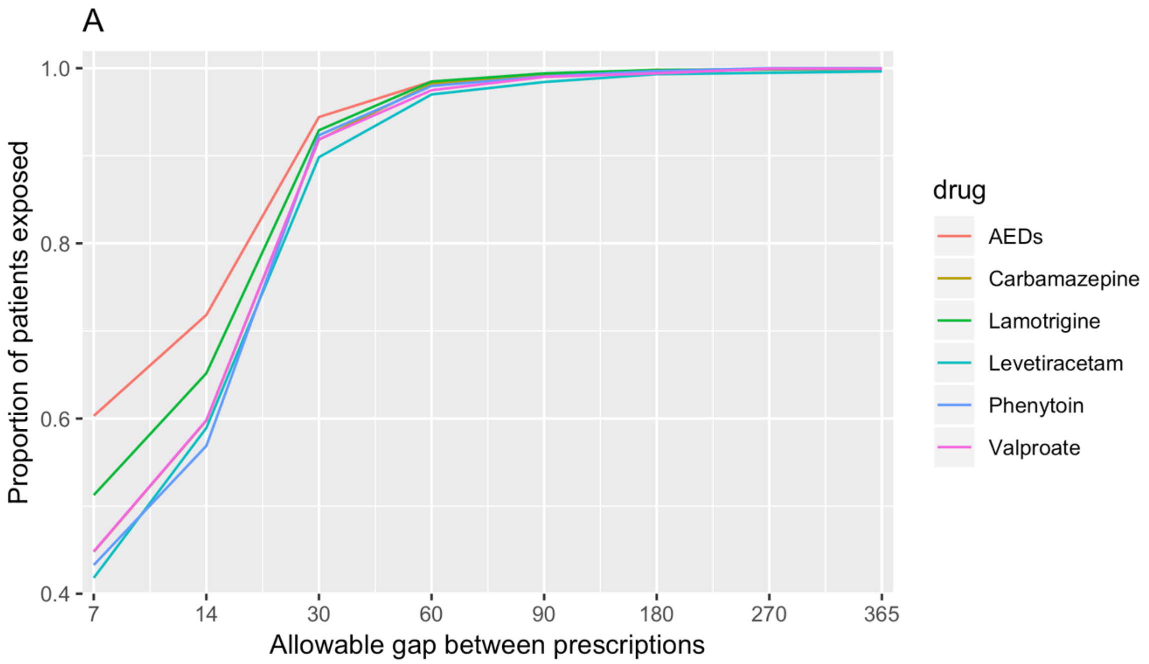
The proportion of patients exposed at time of death based on lookback from death was similar across the investigated drugs (Figure 25). Using time periods of less than 30 days resulted in most patients (70%-80%) being classified as past users but using a timeframe duration of 30 days increased the number of patients classified as current users to between 50 and 60% depending on the drug of interest. Less than 10% of people who died and were ever prescribed each of the drugs were classified as past users when using a 365-day exposure timeframe.



AEDs	Carbamazepine	Lamotrigine	Levetiracetam	Phenytoin	Valproate
1026	299	212	287	156	316

Figure 25 – Proportion of patients classified as current users at time of death based on different timeframes used to define current use, with the total number of patients who died and had at least 1 prescription for each separate drug shown in the table

The proportion of patients exposed based on a definition of two prescriptions within a desired timeframe was consistent across the different drugs of interest except at low allowable gap values, as shown in Figure 26. Over 60% of patients with two or more AED prescriptions were considered exposed when requiring two prescriptions within 7 days of one another. Most patients with two or more prescriptions (>=90% for all drugs) for each drug were considered exposed when allowing for a gap of 30 days or more between prescriptions.

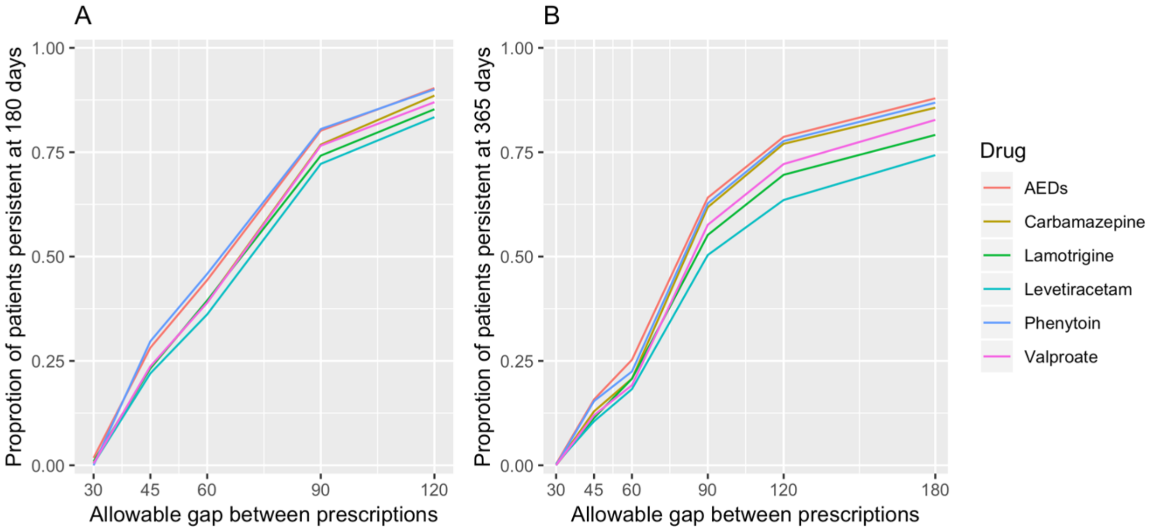


AEDs	Carbamazepine	Lamotrigine	Levetiracetam	Phenytoin	Valproate
5484	1628	1510	1899	640	1709

Figure 26 - Proportion of patients considered exposed to the drug of interest based on a definition of 2 prescriptions within a specified number of days, with the total number of patients who had at least 2 prescriptions for each drug on different dates shown in the table

5.3.3 Persistence

A summary of the proportions of patients who reached 180- and 365-day persistence with different AEDs using different allowable gaps between successive prescriptions can be found in Figure 27 below. Almost no patients were persistent with any of the drugs of interest at either 180 or 365 day after initiation of therapy when using a 30-day allowable gap between prescriptions. The proportion of patients persistent was relatively consistent across drugs, although there was more variation in the proportion of patients persistent at 365 days after initiation with increasing gap length. Levetiracetam had the lowest proportion of patients persistent at 180 and 365 days after initiation.



AEDs	Carbamazepine	Lamotrigine	Levetiracetam	Phenytoin	Valproate
5571	1687	1611	2011	663	1767

Figure 27 - Proportion of patients with at least 1 prescription for AEDs as a class or each of the top 5 AEDs who are persistent at 180 days (A) and 365 days (B) using different allowable gaps between prescriptions, with the total number of patients with at least 1 prescription for each drug shown in the table

In addition to differences across drugs in the proportion of patients persistent at specific time points when using different refill gaps, there were also differences in the median time to first discontinuation. The time to discontinuation was similar across all drugs when using 30-, 45- and 60-day refill gaps. As the allowable gap length increased there was greater variation between the drugs of interest, although AEDs as a class, carbamazepine and phenytoin all had similar results. Levetiracetam had the shortest median time to discontinuation amongst the drugs of interest when using longer allowable gaps.

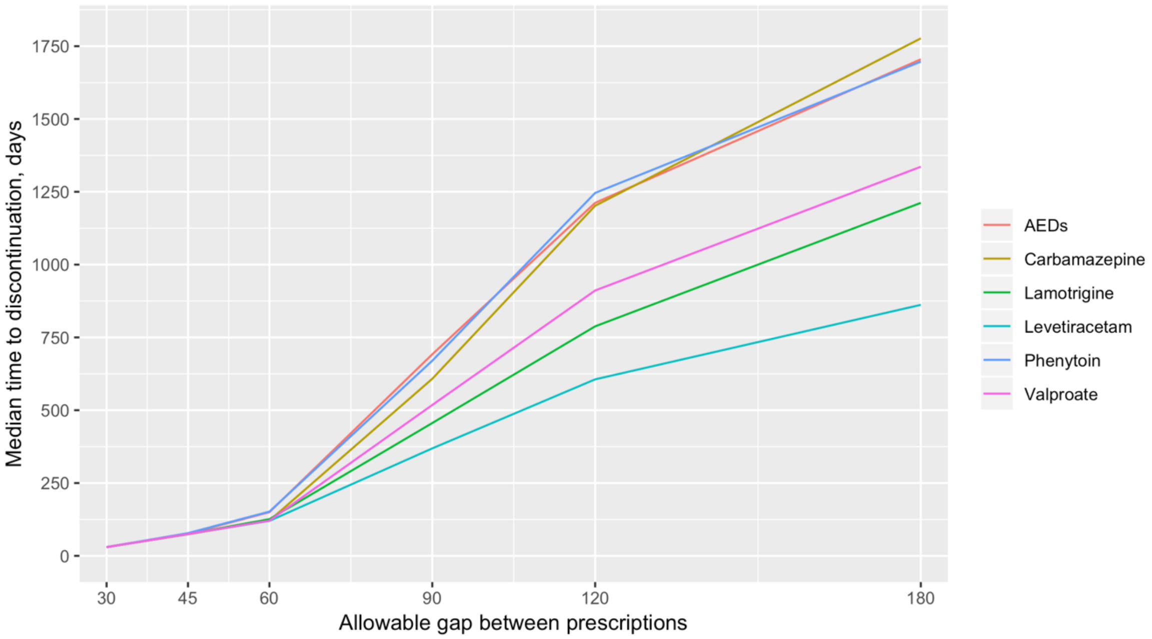


Figure 28 - Summary of the changes in median time to discontinuation with changes in the allowable gap between prescriptions for AEDs as a class and the top 5 AEDs

5.3.4 Effect of exposure definition on observed associations

Of the 5,771 patients in the epilepsy cohort, 2,011 (34.8%) had at least 1 prescription for levetiracetam. Users and non-users had similar gender splits, with more men than women. Levetiracetam users tended to be younger, with a median age of 46 years compared to 52 years among the non-users. Over 50% of the patients in both groups were from the most deprived areas (SIMD quintile 1). A larger proportion of patients in the levetiracetam user group had one or more comorbid conditions than in the non-user group, and a smaller proportion of the levetiracetam users died during follow-up. Median length of follow-up was slightly higher for non-users (4.83 years vs. 4.75 years for users), but over 90% of the users in both groups had a follow-up duration of more than 1 year.

		0 levetiracetam prescriptions N = 3,560	1+ levetiracetam prescriptions N = 2,011
Gender	Female	1,608 (45.2)	931 (46.3)
	Male	1.952 (54.8)	1,080 (53.7)
Age	0 – 9	162 (4.6)	165 (8.2)
	10 – 19	166 (4.7)	182 (9.1)
	20 – 29	244 (6.9)	256 (12.7)
	30 – 39	395 (11.1)	232 (11.5)
	40 – 49	620 (17.4)	301 (15.0)
	50 – 59	645 (18.1)	344 (17.1)
	60 – 69	583 (16.4)	254 (12.6)
	70 – 79	480 (13.5)	172 (8.6)
	80+	265 (7.4)	105 (5.2)
	Median (IQR)	52 (29)	46 (35)
SIMD Quintile	1	1,862 (52.3)	1,017 (50.6)
	2	624 (17.5)	332 (16.5)
	3	408 (11.5)	208 (10.3)
	4	289 (8.1)	192 (9.5)
	5	299 (8.4)	205 (10.2)
	Unknown	78 (2.2)	57 (2.8)
Charlson comorbidity score	0	2,971 (83.5)	1,604 (79.8)
(weighted)	1	307 (8.6)	201 (10.0)
	2	155 (4.4)	76 (3.8)
	3+	140 (3.9)	112 (5.6)
Died	N (%)	739 (20.8)	287 (14.3)
Length of follow-up	Median	4.83	4.75
(years)	IQR	1.43	2.16
	Total	14,360.0	7,629.3
	Range	0.03 – 4.99	0.05 – 4.99
	Patients with > 180d	3,470 (97.5)	1,969 (97.9)
	Patients with >365d	3,351 (94.1)	1,877 (93.3)

Table 28 - Summary of demographic characteristics of cohort patients, split according to whether they ever received at least 1 prescription for levetiracetam

Figure 29, below, shows the hazard ratios and 95% CIs associated with exposure to levetiracetam based on the time-fixed definitions of exposure. There was little variation in the observed hazard ratio based on method. Across all models, the hazard ratios ranged from 1.13-1.14. None of the models showed a statistically significant association between levetiracetam and all-cause mortality (p>0.05). The hazard ratios for age, sex and comorbidity score were consistent across all models.

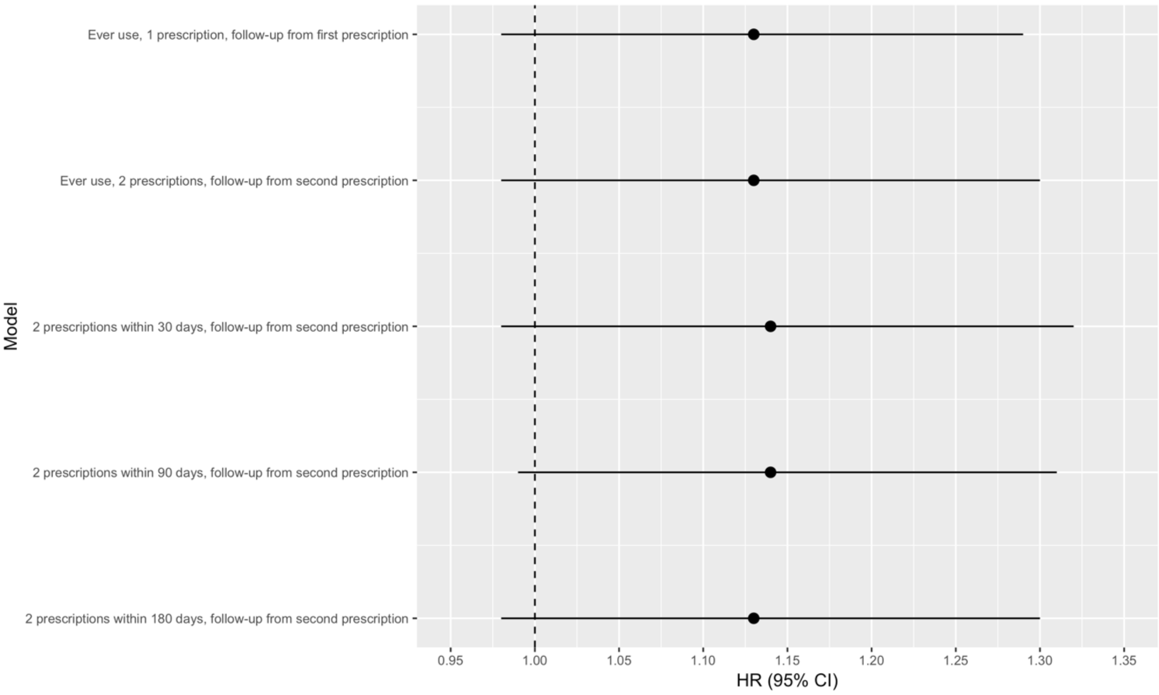


Figure 29 - Association of levetiracetam exposure on all-cause mortality using time-fixed exposure definitions

As shown in Figure 30, there was larger variation in the observed associations when using time-varying definitions of exposure, with hazard ratios between 0.53 and 1.14. In the models where exposure was defined in successive 30-day windows (HR 0.53, 95%CI 0.44 – 0.66) and 90-day windows (HR 0.76, 95%CI 0.65 – 0.89) across follow-up there was a statistically significant reduction in all-cause mortality associated with exposure to levetiracetam ($p<0.001$). In the remaining models, the association between exposure and outcome was not statistically significant ($p>0.01$). As with the models using time-fixed exposure definitions, the observed associations between the other covariates (age, sex and comorbidity score) were consistent across all models.

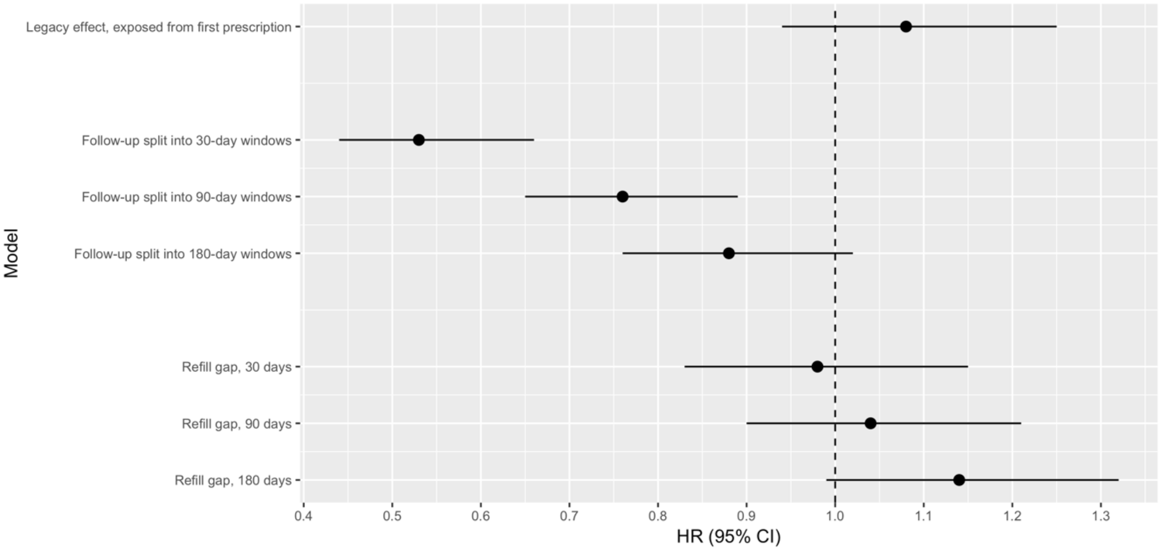


Figure 30 - Association of levetiracetam exposure on all-cause mortality using time-varying exposure definitions

5.4 Discussion

The results described above demonstrate how altering the definition of drug exposure can impact the number of patients classified as exposed to a drug and the duration of treatment episodes and how these can, in turn, affect the association between exposure and outcome.

5.4.1 Effect of exposure definition on the estimated number of exposed patients

5.4.1.1 Ever use

Ever use is the most simplistic definition of exposure, but it is still often used in pharmacoepidemiological research to define cohorts, patient sub-groups and as a proxy for comorbidities or outcomes. In the comparisons of the proportions of patients who met different thresholds, there was variation across different drugs within the same class, and even greater variation across the different classes. The drugs which were used to treat chronic or ongoing conditions such as AEDs, anti-depressants and lipid-regulating drugs retained a higher proportion of patients with increasing thresholds than antibacterials, which are typically used short-term. Analgesics are used for a wider variety of indications and over a wider range of timeframes, which is likely the reason they retain a larger proportion of patients with increasing exposure threshold compared to antibacterials, but a smaller proportion compared to drugs which are only used for chronic conditions. Researchers need to account for these differences when setting exposure thresholds. For example, the same threshold may not be relevant for all drugs if defining a number of comorbidities based on pharmacy data. A threshold value of 1 prescription was the most common inclusion criterion for cohorts or sub-groups across the studies included in the systematic review in Chapter 2, but thresholds from 1-4 prescriptions were reported. In the examples tested here, a threshold of 4 prescriptions would classify 96% of the total users of AEDs as exposed, but only 84% of anti-depressant users, 73% of analgesic users and 58% of antibacterial users. These may be clinically relevant proportions, in cases such as for antibacterial or analgesic drugs they could be a way to select only patients with chronic exposure to these drugs, or certain conditions of interest.

There were also differences observed between the proportions of patients exposed when counting one prescription per day versus all prescriptions for certain drug classes or individual drugs. This should be taken into consideration when cleaning data, as not accounting for this variation will cause systematic differences in the measurement of exposure across the population for different drugs or classes, particularly where the drugs of interest can be prescribed in multiple dosages or formulations at once, as can be the case with AEDs.

5.4.1.2 Use at time

In the results described above there was a larger difference across individual drugs when defining new use at baseline than when defining current use at time of death. This may be due to changes in the patterns of prescribing of certain drugs at population level around the start of follow-up, as both levetiracetam and lamotrigine are newer AEDs and had a smaller proportion of existing users than older drugs or AEDs as a class when longer lookback periods were used. The most commonly used timeframes for defining new use at baseline in the reviewed literature were 6 months and 12 months, but there were examples of studies using windows of varying lengths between 2 months and 3 years.

Defining exposure before an event or categorising patients into current or past use of a drug of interest at the time of an event (for example, death or hospitalisation) was most commonly done using 6- and 12-month windows, but 90- and 120-day windows were also used. In both definitions, the largest differences in the proportions of patients exposed between different windows were seen when using smaller gaps; for AEDs as a class, 64% of patients were considered current users at death using a 30-day window compared to 94% with a 180-day window and 97% with a 365-day window. When defining new use of AEDs at baseline, a 30-day window resulted in 91% of patients being classified as new users, compared to 33% with a 180-day lookback and 31% with a 365-day lookback.

5.4.1.3 Persistence

The variations in the proportion of patients persistent with the drugs investigated at 180- and 365-days were largely similar across the drugs of interest, although there were outliers at each of the gap values. The median time to discontinuation was similar across the different drugs with shorter allowable gap lengths. This likely indicates that these gaps are not long enough to account for the coverage of individual prescriptions for these drugs. Additionally, this may not account for database specific issues with some of the dates recorded within the prescribing data. In cases where the date a prescription was prescribed or dispensed are not available, the date that the pharmacist was reimbursed is imputed, and this is typically the last date of the month. At higher gap values, there was variation across the individual drugs. The median time to discontinuation for newer drugs such as levetiracetam and lamotrigine was lower than the older drugs. As described in Chapter 4, prescribing of these drugs increased over the study period, so it is possible that some discontinuations are due to censoring rather than exceeding the allowable gap. A range of allowable gaps from 14 to 365 days was common across the reviewed literature. In the results described here, shorter refill gaps do not appear to be sufficient to account for the legitimate gaps between prescriptions. In these analyses, no measure of the coverage for each prescription is used, so longer allowable gaps need to be used to account for the supply from each prescription. Both median time to discontinuation and the proportion of patients persistent at 365 days after initiation increased markedly for all drugs when using an allowable gap of 90 days or longer, as these gaps allow for more prescriptions to be included in the individual periods of persistent use. As described previously, allowing for too large a gap between successive prescriptions introduces the potential for error by misclassifying time where patients legitimately had no supply of the drug of interest.

5.4.2 Effect of exposure definition on observed exposure-outcome associations

The results described above highlight the potential for variation in the drug exposure variables generated based on routinely collected data, but it is also important to understand the impact this variation can have on the results of a clinical research study. Although most of the Cox proportional hazards models above did not show a statistically significant association between levetiracetam exposure and all-cause mortality, comparing observed hazard ratios shows the variation that can occur in observed association when only the definition of drug exposure is adjusted.

As described above, the hazard ratios observed across the different time-fixed definitions of exposure were minimal, despite the size of the exposed population being different under each definition. The differences in the observed hazard ratios were larger in the models using time-varying methods of exposure, likely due to the fact that there is more potential for variation from patient to patient as exposure status changes over time. This is particularly true of the methods which split the follow-up into smaller windows. The statistically significant models were the ones with follow-up split into the smallest windows (60 and 90 days respectively) – as with the time-fixed measures above, these were the most specific definitions of exposure used. When splitting follow-up into windows or defining persistent time, use of longer windows increased the hazard ratio observed.

These results show the importance of carefully setting exposure definitions in pharmacoepidemiological research, as incorrectly or imprecisely defining exposure can potentially have a large impact on the results of a study. It is also important for researchers to carefully identify which method has been used to define exposure when reporting to maximise transparency and reproducibility. The models showed a wide range of both positive and negative associations between levetiracetam exposure and all-cause mortality driven exclusively by the changes in the definition of drug exposure. These results would have different clinical implications if reported as the results from a real clinical study. Utilising a time-varying definition of exposure which takes account of the duration of exposure periods will help to minimise the impact of immortal time bias on the results. Sensitivity analysis could be implemented to assess different potential exposure definitions, and even different thresholds within the definitions of interest to identify potential biases in the methods being considered.

5.4.3 Limitations

Although these results highlight that there is a difference in the observed associations depending on exposure definition used, they do not confirm that one method is better than another per se. Researchers should aim to define exposure as precisely as possible based on the data available – it is more realistic to allow for variation in exposure status over time as this reflects reality. However, care needs to be taken to ensure that too many assumptions are not made where the data do not exist to support them. If the data only support a simple definition of exposure, this may still provide more valid results than an ill-defined complex exposure definition, as the assumptions required could result in larger potential for error. It can be useful to use a combination of different methods for different purposes within a study. For example, the study population could be defined as including patients who had at least 1 prescription for a drug of interest to include as many patients as possible, then another method could be used to differentiate between those who had long-term exposure to the drug and those who did not. As discussed in Chapter 2, it is important to specifically report how exposure was defined based on the data available when publishing research findings. This will help increase transparency, giving other researchers a better understanding of the limitations of the available data, how valid the results are and will make it easier to replicate results using other data sources.

As mentioned previously, even if the exposure variable defined is able to perfectly represent the available data, other issues such as missing data can still introduce bias, and routinely collected prescribing data only represent medications which were prescribed or dispensed so there is still no certainty that patients took the medications at all or as instructed.

A limitation of the results is the lack of inclusion of additional methods for constructing treatment episodes based on individual prescriptions. The use of the number of days supplied on individual prescriptions was one of the most commonly used methods for defining exposure amongst the studies included in the systematic review. Assessing differences in durations of treatment episodes across different daily dose/duration methods and comparing current persistence measures to persistence with coverage would be of interest, as it is a more detailed method which has the potential to more accurately capture changes in exposure status over time whilst also providing more potential for bias if incorrectly applied. These methods were not included in the comparisons in this chapter due to the lack of individual level data on prescribed doses within the dataset used. While it is still possible to use the DDD or assumptions based on the number of tablets or units of medication per day, the decision was made to exclude these from this study as they are prone to misrepresent exposure time for drugs, such as AEDs, where there can be variation from patient to patient and over time in the prescribed dose, and therefore the results from these methods would be of limited value without the comparison to individualised episodes.

5.4.4 Further work

There are a number of opportunities for further work on this topic. As stated above, the lack of patient-specific dosage instructions in the data currently available limited the ability to look at the construction of detailed treatment episodes, but these would likely contribute to more precise estimates of the association between exposure and outcome as, if applied correctly, these methods have the capacity to more accurately capture the reality of how long the patient was exposed to the drug and how their exposure status changed over the period of interest. In addition to considering periods of exposure, adherence to treatment within these periods is also of interest as an additional covariate for modelling the risk associated with outcomes related to AED use. Poor adherence to AEDs has been shown to be associated with poor outcomes and increased hospitalisations, with one study demonstrating a threefold increased risk of mortality associated with nonadherence.(270) The outcome investigated in this chapter was used simply to illustrate the potential for variation in observed associations, but doing a similar study using an established exposure outcome relationship would be an even more effective way of highlighting the need for accurate representation of drug exposure in pharmacoepidemiological research.

5.4.5 Summary

Defining patients' exposure to medications of interest is a key step in pharmacoepidemiological research. The results described above show how changing the definition of exposure (either using a different method or adjusting the threshold for exposure) can affect the measured size of the exposed population and the duration of treatment episodes for exposed patients, and in turn how that can affect the observed association between the exposure of interest and an outcome. This highlights the importance of carefully planning data cleaning, utilising sensitivity testing and ensuring that the most precise definition of drug exposure possible is used based on clinically relevant definitions and an understanding of the limitations of the data available in order to minimise exposure misclassification and maximise validity of study results. Additionally, all of these processes must be accurately reported when publishing research in order to maximise transparency and reproducibility of research.

6 Summary and conclusions

6.1 Summary

The aim of this thesis was to develop and validate a set of flexible, reusable functions for generating common drug exposure variables based on routinely collected prescribing data and to demonstrate their utility in clinical research. A range of common methods for quantifying drug exposure were identified and classified based on a systematic review of literature in the field. The information gathered on these methods was used to provide structure to construct an R package, prescribeR, containing functions developed to generate exposure variables from individual prescriptions. This package was then tested by using it to generate the required exposure variables in two example clinical studies using data for a cohort of patients with epilepsy. In the first, the package was used to generate a range of variables in order to investigate the impact of varying drug exposure definition on the observed association between levetiracetam and all-cause mortality in a cohort of patients with epilepsy. In the second, the prescribeR package was used to measure the rates of 1-year persistence to different AEDs in the same cohort of patients.

The main strength of this thesis overall is that it provides an in-depth examination of the role of exposure quantification in pharmacoepidemiological research using routinely collected data. The studies identified and reviewed in the Chapter 2 highlighted the variety of research being published based on routinely collected data, and also allowed for an examination of the potential strengths and limitations of the different methods being used to quantify exposure across these different studies. When selecting a method of quantifying drug exposure, it is important to consider the potential for both exposure misclassification and time-related bias.

Minimising exposure misclassification requires an understanding of which methods are suitable for the research question being posed, as well as the structure and content of the database being used and the way the data are collected and potential for missing or incomplete data. It is important to balance defining exposure as precisely as possible in clinical research with minimising the potential for error from making assumptions without the data to support them. For example, if the datasets being used do not contain information regarding patient specific dosage instructions, assuming the duration of prescriptions for certain drugs based on other measures may impact the accuracy of the exposure measurement. On the other hand, too broad a definition of exposure will introduce error by conflating different levels of exposure. It is also important to consider the study design in order to ensure that the time period for defining exposure is correctly set out in order to minimise the risk of immortal time bias.

The classes of exposure quantification methods defined in the systematic review provided a clear outline for the development of the prescribeR package, which contains functions for ever use, use at time point, daily dose and persistence methods alongside functions for standardising data and generating data summaries. The benefit of using a package such as prescribeR in clinical research is that it provides a standardised set of tools for easily quantifying drug exposure, as well as a framework for clear reporting on how the data were processed. This, in turn, makes it easier for the results to be reproduced and validated with other data.

As shown in Chapter 5, altering the method of defining exposure without changing any of the other study parameters can have a large impact on the observed associations between drug exposure and an outcome, so ensuring the variables generated represent the data available as accurately as possible is essential to ensuring that the results of the study are valid. In addition to minimising the risk of exposure misclassification within individual studies, it is important to be able to reproduce and validate the results obtained using data from other sources in order to ensure that database-specific limitations aren't impacting the results. For this reason, it is important that published studies make sure to clearly describe the processes used to prepare the data, not just the statistical analyses.

A paper summarising the conclusions from two studies published by a joint International Society of Pharmacoepidemiology and International Society for Pharmacoeconomics and Outcomes Research (ISPE/ISPOR) task force on real-world evidence in healthcare decision making concluded that increased transparency in the planning and reporting of studies using routinely collected data was an essential component in ensuring that the results of these types of studies had the appropriate impact on healthcare decision making. The STROBE, RECORD and RECORD-PE statements provide checklists for authors detailing items which should be included in studies reporting on observational studies, with RECORD focusing specifically on the reporting of studies using routinely collected data.(221, 271) The RECORD checklist recommends that codes and algorithms used to specify different variables in the study should be provided where possible. As discussed in Chapter 3, one of the primary benefits of using a package such as prescribeR during the data preparation process is that it standardises parts of the process, providing a structure to report the way the data have been used, as demonstrated in the methods section of Chapter 6. This then allows other researchers to understand at a glance what processes were applied to generate the results and then to apply the same methods to generate the variables of interest from different datasets.

6.2 Limitations

Limitations of the individual components of this thesis are discussed in the relevant chapters, but there are also limitations which apply to the thesis as a whole. The primary focus of this thesis was the methods used to quantify drug exposure in studies using routinely collected healthcare data. Although this is an essential component of preparing routinely collected data for use in research, there are a number of other processes involved in cleaning and enriching data. A number of these issues were identified throughout the thesis, and attempts were made to address them when preparing the data used in the analyses described in Chapters 4 and 5, but the potential effect that altering the decisions made or the impact that errors in the final datasets can have on study results was not assessed in this thesis. One study which investigated the impact of decisions made to clean missing or implausible values in the quantity dispensed, dosage instructions and stop dates for individual prescriptions showed that there was variation in the observed association between glucocorticoids or oral hypoglycaemic drugs and cardiovascular events based on the definitions used.(272)

Similarly, a number of limitations related to the way that routinely collected data are handled were identified in this thesis, but again the impact of these issues on study results was not assessed in depth here. Although PIS contains most of the prescriptions prescribed and dispensed in Scotland, it does not account for medications administered or dispensed within hospitals, which may impact studies where the drug of interest is often prescribed during hospitalisations or on discharge. As described previously in Chapter 1, PIS does not contain data regarding diagnoses or indication for prescribing. Diagnoses can be inferred from other data such as hospitalisations or primary care data, but the completeness and accuracy of these data should also be considered. Where these data sources are not available, surrogate markers can be used, as was the case in the definition of epilepsy in this thesis. Without additional data, it is not possible to quantify the impact of these assumptions on case definition and study results. Finally, the prescribeR package was developed and tested using prescribing data from a single source (PIS), and the development of the processes used within the functions were based on experience using this data. Therefore, there may be steps required in handling data from other sources which are not currently included in the prescribeR functions. However, one of the benefits of hosting the package on GitHub is that it is possible for users to report issues or suggest improvements to the package, providing the opportunity to resolve any issues or limitations that are raised and therefore improve the package's utility.

6.3 Future work

There are a number of opportunities for further work based on the research described in this thesis. At present, the prescribeR package is primarily focused on the generation of individual-level drug exposure variables based on the common methods identified in the systematic review in Chapter 2, but it is not necessarily comprehensive. The structure of packages in R means it is easy to iterate on the current version to include expansions to existing functions or to add new functions as required. Potential additions to the package could include additional functions for examining population-level prescribing trends or functions for constructing variables based on more complex prescribing behaviours such as co-prescribing, polypharmacy and medication switching. Beyond exposure quantification, functions could be added for other common data preparation or data enrichment tasks, such as adding variables from other datasets or standardising field syntax.

There are also a number of opportunities for further work assessing the impact of varying drug exposure definition on the results obtained in clinical studies. As stated in Chapter 5, the lack of patient-specific dosage instructions in the data currently available limited the ability to compare different methods of generating individual treatment episodes. Treatment episodes based on patient-specific dosage information would likely provide more precise estimates of the association between exposure and outcome, as these methods have the capacity to more accurately capture the reality of how long the patient was exposed to the drug and how this changed over the period of interest. Additionally, the outcome investigated in this thesis was used simply to illustrate the potential for variation in observed associations, but a similar study using an established, clinically relevant exposure outcome relationship would be an even more effective way of highlighting the need for accurate representation of drug exposure in pharmacoepidemiological research.

6.4 Final conclusions

Routinely collected data have the potential to be a vital resource for pharmacoepidemiology research and, in turn, for generating evidence for clinical decision making. However, the data are not without limitations, so it is important that researchers address these during data preparation and analysis. Quantifying drug exposure is an essential step in research using routinely collected data, as bias introduced in this step has an impact on the results obtained and therefore impacts the validity of the evidence generated. The prescribeR package provides researchers with a set of straightforward, reusable functions for generating drug exposure variables using a number of common methods, as well as a simple structure for clearly reporting the method used.

Appendix 1 - List of anti-epileptic drugs

Drug Name
Brivaracetam
Carbamazepine
Clobazam
Clonazepam
Eslicarbazepine
Ethosuximide
Gabapentin
Lacosamide
Lamotrigine
Levetiracetam
Oxcarbazepine
Perampanel
Phenobarbital
Phenytoin
Pregabalin
Primidone
Retigabine
Rufinamide
Sodium valproate
Stiripentol
Tiagabine
Topiramate
Vigabatrin
Zonisamide

List of References

1. Smith P, Morrow R, Ross D, editors. Field Trials of Health Interventions: A Toolbox. 3rd ed. Oxford, UK: OUP Oxford; 2015.
2. Eurostat. Medicine use statistics 2017 [Available from: http://ec.europa.eu/eurostat/statistics-explained/index.php/Medicine_use_statistics, access date 20/11/2017.
3. Qato DM, Wilder J, Schumm P, Gillet V, Alexander GC. Changes in Prescription and Over-the-Counter Medication and Dietary Supplement Use Among Older Adults in the United States, 2005 vs 2011. Jama Internal Medicine. 2016;176(4):473-82.
4. NHS Digital. Areas of Interest - Prescribing 2020 [Available from: https://digital.nhs.uk/data-and-information/areas-of-interest/prescribing, last access date 13/06/2020.
5. NHS Digital. Prescriptions dispensed in the community, England 2006 - 2016 England2017 [Available from: https://files.digital.nhs.uk/publication/s/o/pres-disp-com-eng-2006-16-rep.pdf, last accessed 13/06/2020.
6. Jones DS, Podolsky SH. The history and fate of the gold standard. Lancet. 2015;385(9977):1502-3.
7. Sertkaya A, Wong H-H, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. Clinical Trials. 2016;13(2):117-26.
8. Silverman SL. From Randomized Controlled Trials to Observational Studies. The American Journal of Medicine. 2009;122(2):114-20.
9. Blanco C, Olfson M, Goodwin RD, Ogburn E, Liebowitz MR, Nunes EV, et al. Generalizability of clinical trial results for major depression to community samples: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. Journal of Clinical Psychiatry. 2008;69(8):1276-80.
10. Treweek S, Dryden R, McCowan C, Harrow A, Thompson AM. Do participants in adjuvant breast cancer trials reflect the breast cancer patient population? European Journal of Cancer. 2015;51(8):907-14.
11. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015;16.
12. Frieden TR. Evidence for Health Decision Making - Beyond Randomized, Controlled Trials. New England Journal of Medicine. 2017;377(5):465-75.
13. International Society for Pharmacoepidemiology. About Pharmacoepidemiology [Available from: https://www.pharmacoepi.org/about-ispe/overview/, last accessed 13/06/2020.
14. World Health Organisation. WHO | Pharmacovigilance 2015 [updated 2015-11-20 22:46:20. Available from: https://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/, last accessed 13/06/2020.
15. Melchert M, List A. The thalidomide saga. The International Journal of Biochemistry & Cell Biology. 2007;39(7):1489-99.
16. Stolberg SG. Heart Drug Withdrawn as Evidence Shows It Could Be Lethal: The New York Times; 1998 [updated 19980609. Available from: https://www.nytimes.com/1998/06/09/us/heart-drug-withdrawn-as-evidence-shows-it-could-be-lethal.html, last accessed 13/06/2020.
17. Pollack A. Abbott Labs Withdraws Meridia From Market: The New York Times; 2010 [updated 20101008. Available from:

https://www.nytimes.com/2010/10/09/health/09drug.html, last accessed 13/06/2020.
18. Onakpoya IJ, Heneghan CJ, Aronson JK. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. Bmc Medicine. 2016;14.
19. Hennessy S. Use of health care databases in pharmacoepidemiology. Basic & Clinical Pharmacology & Toxicology. 2006;98(3):311-3.
20. Murphy M, Alavi K, Maykel J. Working with Existing Databases. Clinics in Colon and Rectal Surgery. 2013;26(1):5-11.
21. van Trijffel E, Oostendorp RAB, Elvers JWH. Routinely collected data as real-world evidence for physiotherapy practice. Physiotherapy Theory and Practice. 2019;35(9):805-9.
22. NHS Research Scotland. Our population | NHS Research Scotland 2020 [Available from: https://www.nhsresearchscotland.org.uk/research-in-scotland/data/our-population, last accessed 13/06/2020.
23. Information Services Division Scotland. Community Health Index Number - Data Dictionary [Available from: https://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?ID=128&Title=CHI%20Number, last accessed 13/06/2020.
24. Scottish Government. The use of the CHI (Community Health Index) to support integrated care across the NHS in Scotland Scotland: Scottish Government eHealth Division; 2013 [1.1:[Available from: https://www.ehealth.scot/resources/information-governance/publications/, last accessed 13/06/2020.
25. ISD Scotland. Prescribing Information System | National Data Catalogue | National Datasets | ISD Scotland | Information Services Division 2016 [Available from: https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=9, last accessed 13/06/2020, last accessed 13/06/2020.
26. Alvarez-Madrazo S, McTaggart S, Nangle C, Nicholson E, Bennie M. Data Resource Profile: The Scottish National Prescribing Information System (PIS). International Journal of Epidemiology. 2016;45(3):714–5f.
27. Open Risk Manual. Data Completeness - Definition 2016 [Available from: https://www.openriskmanual.org/wiki/Data_Completeness#cite_ref-1, last accessed 13/06/2020.
28. Open Risk Manual. Data Accuracy - Definition 2016 [Available from: https://www.openriskmanual.org/wiki/Data_Accuracy, last accessed 13/06/2020.
29. Takahashi Y, Nishida Y, Asai S. Utilization of health care databases for pharmacoepidemiology. Eur J Clin Pharmacol. 2012;68(2):123-9.
30. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding Control in Healthcare Database Research Challenges and Potential Approaches. Medical Care. 2010;48(6):S114-S20.
31. Colantonio LD, Kent ST, Kilgore ML, Delzell E, Curtis JR, Howard G, et al. Agreement between Medicare pharmacy claims, self-report, and medication inventory for assessing lipid-lowering medication use. Pharmacoepidemiology and Drug Safety. 2016;25(7):827-35.
32. Curtis JR, O. WA, Jeroan A, Allison F, H. KS, G. SK. Agreement and validity of pharmacy data versus self-report for use of osteoporosis medications among chronic glucocorticoid users. Pharmacoepidemiology and Drug Safety. 2006;15(10):710--8.
33. Porta M. A Dictionary of Epidemiology. 5 ed: Oxford University Press; 2008.
34. Harpe SE. Using Secondary Data Sources for Pharmacoepidemiology and Outcomes Research. Pharmacotherapy. 2009;29(2):138-53.

35. Wettermark B, Zoega H, Furu K, Korhonen M, Hallas J, Norgaard M, et al. The Nordic prescription databases as a resource for pharmacoepidemiological research--a literature review. Pharmacoepidemiol Drug Saf. 2013;22(7):691-9.

36. Committee JF. British National Formulary, London: BMJ Group and Pharmaceutical Press; 2017 [June 2017:[Available from: https://www.medicinescomplete.com/mc/bnflegacy/current/, last accessed 13/06/2020.

37. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ-British Medical Journal. 2009;339(b2535).

38. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet. 2012;380(9836):37-43.

39. World Health Organisation. WHOCC - Definition and general considerations 2016 [Available from: https://www.whocc.no/ddd/definition_and_general_considera/, last accessed 13/06/2020.

40. Cramer JA, Roy A, Burrell A, Fairchild CJ, Fuldeore MJ, Ollendorf DA, et al. Medication compliance and persistence: Terminology and definitions. Value in Health. 2008;11(1):44-7.

41. Osterberg L, Blaschke T. Drug therapy - Adherence to medication. New England Journal of Medicine. 2005;353(5):487-97.

42. Raebel MA, Schmittdiel J, Karter AJ, Konieczny JL, Steiner JF. Standardizing Terminology and Definitions of Medication Adherence and Persistence in Research Employing Electronic Databases. Medical Care. 2013;51(8):S11-S21.

43. Dregan A, Charlton J, Wolfe CDA, Gulliford MC, Markus HS. Is Sodium Valproate, an HDAC inhibitor, associated with reduced risk of stroke and myocardial infarction? A nested case-control study. Pharmacoepidemiology and Drug Safety. 2014;23(7):759-67.

44. Rushton CA, Stromberg A, Jaarsma T, Kadam UT. Multidrug and optimal heart failure therapy prescribing in older general practice populations: A clinical data linkage study. BMJ Open. 2014;4(1).

45. Winterstein AG, Castillo JMH, Xu D, Liu W, Antonelli PJ. Ototopical neomycin exposure in children with nonintact tympanic membranes. Laryngoscope. 2012;122(11):2529-32.

46. Etminan M, Skeldon SC, Goldenberg SL, Carleton B, Brophy JM. Testosterone therapy and risk of myocardial infarction: a pharmacoepidemiologic study. Pharmacotherapy:The Journal of Human Pharmacology & Drug Therapy. 2015;35(1):72-8.

47. Suehs BT, Davis C, Galaznik A, Joshi AV, Zou KH, Patel NC. Association of out-of-pocket pharmacy costs with adherence to varenicline. Journal of Managed Care Pharmacy. 2014;20(6):592-600.

48. Momen NC, Olsen J, Gissler M, Kieler H, Haglund B, Li J. Exposure to systemic antibacterial medications during pregnancy and risk of childhood cancer. Pharmacoepidemiology and Drug Safety. 2015;24(8):821-9.

49. Lee MC, Lee CH, Shu CC, Pong WB, Lan CC, Wang JY, et al. The impact of diabetes mellitus and its control on the development of tuberculosis: A nationwide longitudinal study in Taiwan. Pharmacoepidemiology and Drug Safety. 2013;22(9):995-1003.

50. Driessen JHM, van Onzenoort HAW, Starup-Linde J, Henry R, Neef C, van den Bergh J, et al. Use of dipeptidyl peptidase 4 inhibitors and fracture risk

compared to use of other anti-hyperglycemic drugs. Pharmacoepidemiology and Drug Safety. 2015;24(10):1017-25.

51. Braunstein D, Hardy A, Boucherie Q, Frauger E, Blin O, Gentile G, et al. Antidepressant adherence patterns in older patients: use of a clustering method on a prescription database. Fundamental and Clinical Pharmacology. 2017;31(2):226-36.

52. Sultana J, Italiano D, Spina E, Cricelli C, Lapi F, Pecchioli S, et al. Changes in the prescribing pattern of antidepressant drugs in elderly patients: An Italian, nationwide, population-based study. European Journal of Clinical Pharmacology. 2014;70(4):469-78.

53. Romanelli RJ, Jukes T, Segal JB. Compliance after switching from branded to generic statins. Pharmacoepidemiology and drug safety. 2014;23(10):1093-100.

54. Shore S, Carey EP, Turakhia MP, Jackevicius CA, Cunningham F, Pilote L, et al. Adherence to dabigatran therapy and longitudinal patient outcomes: Insights from the Veterans Health Administration. American Heart Journal. 2014;167(6):810-7.

55. Pottegard A, Friis S, Andersen M, Hallas J. Use of benzodiazepines or benzodiazepine related drugs and the risk of cancer: A population-based case-control study. British Journal of Clinical Pharmacology. 2013;75(5):1356-64.

56. Adams AL, Black MH, Zhang JL, Shi JM, Jacobsen SJ. Proton-pump inhibitor use and hip fractures in men: A population-based case-control study. Annals of Epidemiology. 2014;24(4):286-90.

57. Taitel M, Fensterheim L, Kirkham H, Sekula R, Duncan I. Medication days' supply, Adherence, wastage, and cost among chronic patients in Medicaid. Medicare and Medicaid Research Review. 2012;2(3):E1-E13.

58. Czaja AS, Valuck RJ, Anderson HD. Comparative safety of selective serotonin reuptake inhibitors among pediatric users with respect to adverse cardiac events. Pharmacoepidemiology and Drug Safety. 2013;22(6):607-14.

59. Carter CT, Changolkar AK, Scott McKenzie R. Adalimumab, etanercept, and infliximab utilization patterns and drug costs among rheumatoid arthritis patients. Journal of Medical Economics. 2012;15(2):332-9.

60. Zeng F, Knoth RL, Patel BV, Kim E, Tran QV, Jing Y. Impact of health plan restrictions on antipsychotic medication adherence and persistence. American Journal of Pharmacy Benefits. 2012;4(1):e22-e31.

61. Birt J, Johnston J, Nelson D. Exploration of claims-based utilization measures for detecting potential nonmedical use of prescription drugs. Journal of Managed Care Pharmacy. 2014;20(6):639-46.

62. Campbell JH, Schwartz GF, Labounty B, Kowalski JW, Patel VD. Patient adherence and persistence with topical ocular hypotensive therapy in real-world practice: A comparison of bimatoprost 0.01% and travoprost Z 0.004% ophthalmic solutions. Clinical Ophthalmology. 2014;8(pp 927-935).

63. Souverein PC, Koster ES, Colice G, van Ganse E, Chisholm A, Price D, et al. Inhaled Corticosteroid Adherence Patterns in a Longitudinal Asthma Cohort. Journal of Allergy and Clinical Immunology: In Practice. 2016;5(2):448-56.e2.

64. Ratanawongsa N, Karter AJ, Quan J, Parker MM, Handley M, Sarkar U, et al. Reach and validity of an objective medication adherence measure among safety net health plan members with diabetes: A cross-sectional study. Journal of Managed Care and Specialty Pharmacy. 2015;21(8):688-98.

65. Lauffenburger JC, Mayer CL, Hawke RL, Brouwer KLR, Fried MW, Farley JF. Medication use and medical comorbidity in patients with chronic hepatitis C from a US commercial claims database: High utilization of drugs with interaction potential. European Journal of Gastroenterology and Hepatology. 2014;26(10):1073-82.

66. Rasmussen LD, Obel D, Kronborg G, Larsen CS, Pedersen C, Gerstoft J, et al. Utilization of psychotropic drugs prescribed to persons with and without HIV infection: A Danish nationwide population-based cohort study. HIV Medicine. 2014;15(8):458-69.
67. Belleudi V, Di Martino M, Cascini S, Kirchmayer U, Pistelli R, Formoso G, et al. The impact of adherence to inhaled drugs on 5-year survival in COPD patients: a time dependent approach. Pharmacoepidemiology and Drug Safety. 2016;25(11):1295-304.
68. Berard A, Sheehy O. The Quebec Pregnancy Cohort--prevalence of medication use during gestation and pregnancy outcomes. PLoS ONE [Electronic Resource]. 2014;9(4):e93870.
69. Dawson J, Fulton R, McInnes GT, Morton R, Morrison D, Padmanabhan S, et al. Acetaminophen use and change in blood pressure in a hypertensive population. Journal of Hypertension. 2013;31(7):1485-90.
70. Koopmans PC, Bos JH, De Jong van den Berg LT. Are antibiotics related to oral combination contraceptive failures in the Netherlands? A case-crossover study. Pharmacoepidemiology and Drug Safety. 2012;21(8):865-71.
71. Zhang T, Kingwell E, De Jong HJ, Zhu F, Zhao Y, Carruthers R, et al. Association between the use of selective serotonin reuptake inhibitors and multiple sclerosis disability progression. Pharmacoepidemiology and Drug Safety. 2016;25(10):1150-9.
72. Karasneh RA, Murray LJ, Hughes CM, Cardwell CR. Digoxin use after diagnosis of prostate cancer and survival: a population-based cohort study. Pharmacoepidemiology and Drug Safety. 2016;25(9):1099-103.
73. Mc Menamin UC, Murray LJ, Hughes CM, Cardwell CR. Metformin use and survival after colorectal cancer: A population-based cohort study. International Journal of Cancer. 2016;138(2):369-79.
74. Mc Menamin C, Murray LJ, Hughes CM, Cardwell CR. Statin use and breast cancer survival: A nationwide cohort study in Scotland. BMC Cancer. 2016;16(1).
75. Boudreau DM, Wirtz H, Von Korff M, Catz SL, St.John J, Stang PE. A survey of adult awareness and use of medicine containing acetaminophen. Pharmacoepidemiology and Drug Safety. 2013;22(3):229-40.
76. Chui CSL, Man KKC, Cheng CL, Chan EW, Lau WCY, Cheng VCC, et al. An investigation of the potential association between retinal detachment and oral fluoroquinolones: A self-controlled case series study. Journal of Antimicrobial Chemotherapy. 2014;69(9):2563-7.
77. de Jong HJI, Vandebriel RJ, Saldi SRF, van Dijk L, van Loveren H, Cohen Tervaert JW, et al. Angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers and the risk of developing rheumatoid arthritis in antihypertensive drug users. Pharmacoepidemiology and Drug Safety. 2012;21(8):835-43.
78. Bakken MS, Schjott J, Engeland A, Engesaeter LB, Ruths S. Antipsychotic Drugs and Risk of Hip Fracture in People Aged 60 and Older in Norway. Journal of the American Geriatrics Society. 2016;64(6):1203-9.
79. Gomm W, Von Holt K, Thome F, Broich K, Maier W, Fink A, et al. Association of proton pump inhibitors with risk of dementia: A pharmacoepidemiological claims data analysis. JAMA Neurology. 2016;73(4):410-6.
80. Sohn M, Talbert J, Blumenschein K, Moga DC. Atypical antipsychotic initiation and the risk of type II diabetes in children and adolescents. Pharmacoepidemiology & Drug Safety. 2015;24(6):583-91.

81. Ramcharran D, Qiu H, Schuemie MJ, Ryan PB. Atypical antipsychotics and the risk of falls and fractures among older adults: An emulation analysis and an evaluation of additional confounding control strategies. Journal of Clinical Psychopharmacology. 2017;37(2):162-8.
82. Torp-Pedersen A, Jimenez-Solem E, Cejvanovic V, Poulsen HE, Andersen JT. Birth outcomes after exposure to mebendazole and pyrvinium during pregnancy-A Danish nationwide cohort study. Journal of Obstetrics and Gynaecology. 2016;36(8):1020-5.
83. Spoendlin J, Meier C, Jick SS, Meier CR. Bisphosphonate therapy start may transiently increase the risk of tendon rupture in patients with glucocorticoid co-medication: a population-based observational study. Pharmacoepidemiology and Drug Safety. 2016;25(10):1116-23.
84. Shin J, Hunt CM, Suzuki A, Papay JI, Beach KJ, Cheetham TC. Characterizing phenotypes and outcomes of drug-associated liver injury using electronic medical record data. Pharmacoepidemiology and Drug Safety. 2013;22(2):190-8.
85. Pottegard A, Meegaard PM, Holck LHV, Christensen RD, Madsen H, Hallas J. Concurrent use of tramadol and oral vitamin K antagonists and the risk of excessive anticoagulation: A register-based nested case-control study. European Journal of Clinical Pharmacology. 2013;69(3):641-6.
86. Farsani SF, Souverein PC, Van Der Vorst MMJ, Mantel-Teeuwisse AK, Knibbe CAJ, De Boer A. Disease history and medication use as risk factors for the clinical manifestation of type 1 diabetes in children and young adults: An explorative case control study. PLoS ONE. 2014;9(2).
87. Hsu CL, Chang CH, Lin JW, Wu LC, Chuang LM, Lai MS. Histamine-2 receptor antagonists and risk of lung cancer in diabetic patients - an exploratory analysis. Pharmacoepidemiology and Drug Safety. 2013;22(6):632-40.
88. Colhoun HM, Livingstone SJ, Looker HC, Morris AD, Wild SH, Lindsay RS, et al. Hospitalised hip fracture risk with rosiglitazone and pioglitazone use compared with other glucose-Lowering drugs. Diabetologia. 2012;55(11):2929-37.
89. de Jonge L, Zetstra-van der Woude PA, Bos HJ, de Jong-van den Berg LT, Bakker MK. Identifying associations between maternal medication use and birth defects using a case-population approach: an exploratory study on signal detection. Drug Safety. 2013;36(11):1069-78.
90. Gnjidic D, Hilmer SN, Hartikainen S, Tolppanen AM, Taipale H, Koponen M, et al. Impact of high risk drug use on hospitalization and mortality in older people with and without Alzheimer's disease: A national population cohort study. PLoS ONE. 2014;9(1).
91. ter Horst PG, Bos HJ, de Jong-van de Berg LT, Wilffert B. In utero exposure to antidepressants and the use of drugs for pulmonary diseases in children. European Journal of Clinical Pharmacology. 2013;69(3):541-7.
92. Bakken MS, Engeland A, Engesaeter LB, Ranhoff AH, Hunskaar S, Ruths S. Increased risk of hip fracture among older people using antidepressant drugs: Data from the Norwegian Prescription Database and the Norwegian Hip Fracture Registry. Age and Ageing. 2013;42(4):514-20.
93. Norgaard M, Ehrenstein V, Nielsen RB, Bakketeig LS, Sorensen HT. Maternal use of antibiotics, hospitalisation for infection during pregnancy, and risk of childhood epilepsy: A population-based cohort study. PLoS ONE. 2012;7(1).
94. Kim JY, Lee J, Ko YJ, Shin JY, Jung SY, Choi NK, et al. Multi-indication carbamazepine and the risk of severe cutaneous adverse drug reactions in korean elderly patients: A korean health insurance data-based study. PLoS ONE. 2013;8(12).

95. Scherrer JF, Salas J, Bucholz KK, Schneider FD, Burroughs T, Copeland LA, et al. New depression diagnosis following prescription of codeine, hydrocodone or oxycodone. Pharmacoepidemiology and Drug Safety. 2016;25(5):560-8.

96. de Souza AS, Jr., Dos Santos DB, Rey LC, Medeiros MG, Vieira MG, Coelho HL. Off-label use and harmful potential of drugs in a NICU in Brazil: A descriptive study. BMC Pediatrics. 2016;16:13.

97. Neumann A, Weill A, Ricordeau P, Fagot JP, Alla F, Allemand H. Pioglitazone and risk of bladder cancer among diabetic patients in France: A population-based cohort study. Diabetologia. 2012;55(7):1953-62.

98. Gallagher AM, Van Staa TP, Murray-Thomas T, Schoof N, Clemens A, Ackermann D, et al. Population-based cohort study of warfarin-treated patients with atrial fibrillation: Incidence of cardiovascular and bleeding outcomes. BMJ Open. 2014;4(1).

99. Mao Y, Pedersen LH, Christensen J, Vestergaard M, Zhou W, Olsen J, et al. Prenatal exposure to antidepressants and risk of epilepsy in childhood. Pharmacoepidemiology and Drug Safety. 2016;25(11):1320-30.

100. Poluzzi E, Raschi E, Godman B, Koci A, Moretti U, Kalaba M, et al. Pro-arrhythmic potential of oral antihistamines (H1): combining adverse event reports with drug utilization data across Europe. PLoS ONE [Electronic Resource]. 2015;10(3):e0119551.

101. Chang CM, Wu KY, Chiu YW, Wu HT, Tsai YT, Chau YL, et al. Psychotropic drugs and risk of burn injury in individuals with mental illness: a 10-year population-based case-control study. Pharmacoepidemiology and Drug Safety. 2016;25(8):918-27.

102. Nishtala PS, Chyou TY. Real-world risk of diabetes with antipsychotic use in older New Zealanders: a case-crossover study. European Journal of Clinical Pharmacology. 2017;73(2):233-9.

103. Morales DR, Dreischulte T, Lipworth BJ, Donnan PT, Jackson C, Guthrie B. Respiratory effect of beta-blocker eye drops in asthma: population-based study and meta-analysis of clinical trials. British Journal of Clinical Pharmacology. 2016;82(3):814-22.

104. Choi HJ, Park C, Lee YK, Ha YC, Jang S, Shin CS. Risk of fractures in subjects with antihypertensive medications: A nationwide claim study. International Journal of Cardiology. 2015;184(pp 62-67).

105. Bakken MS, Engeland A, Engesaeter LB, Ranhoff AH, Hunskaar S, Ruths S. Risk of hip fracture among older people using anxiolytic and hypnotic drugs: A nationwide prospective cohort study. European Journal of Clinical Pharmacology. 2014;70(7):873-80.

106. Tagalakis V, Tamim H, Blostein M, Hanley JA, Kahn SR. Risk of prostate cancer death in long-Term users of warfarin: A population-based case-control study. Cancer Causes and Control. 2013;24(6):1079-85.

107. Eftekhari K, Ghodasra DH, Haynes K, Chen J, Kempen JH, Vanderbeek BL. Risk of retinal tear or detachment with oral fluoroquinolone use: A cohort study. Pharmacoepidemiology and Drug Safety. 2014;23(7):745-52.

108. Fardet L, Nazareth I, Whitaker HJ, Petersen I. Severe neuropsychiatric outcomes following discontinuation of long-term glucocorticoid therapy: A cohort Study. Journal of Clinical Psychiatry. 2013;74(4):e281-e6.

109. Avgil Tsadok M, Jackevicius CA, Rahme E, Humphries KH, Behlouli H, Pilote L. Sex differences in stroke risk among older patients with recently diagnosed atrial fibrillation. JAMA. 2012;307(18):1952-8.

110. Haukka J, Niskanen L, Partonen T, Lonnqvist J, Tiihonen J. Statin usage and all-cause and disease-specific mortality in a nationwide study. Pharmacoepidemiology and Drug Safety. 2012;21(1):61-9.
111. Jespersen CG, Norgaard M, Friis S, Skriver C, Borre M. Statin use and risk of prostate cancer: A Danish population-based case-control study, 1997-2010. Cancer Epidemiology. 2014;38(1):42-7.
112. Nakafero G, Sanders RD, Nguyen-Van-Tam JS, Myles PR. The association between benzodiazepines and influenza-like illness-related pneumonia and mortality: a survival analysis using UK Primary Care data. Pharmacoepidemiology and Drug Safety. 2016;25(11):1263-73.
113. Yokoyama K, Yamazaki K, Katafuchi M, Ferchichi S. A Retrospective Claims Database Study on Drug Utilization in Japanese Patients with Crohn's Disease Treated with Adalimumab or Infliximab. Advances in Therapy. 2016;33(11):1947-63.
114. Krivoy A, Balicer RD, Feldman B, Hoshen M, Zalsman G, Weizman A, et al. Adherence to Antidepressants Is Associated With Lower Mortality: A 4-Year Population-Based Cohort Study. Journal of Clinical Psychiatry. 2016;77(5):e566-e72.
115. Disantostefano RL, Yeakey AM, Raphiou I, Stempel DA. An evaluation of asthma medication utilization for risk evaluation and mitigation strategies (REMS) in the United States: 2005-2011. Journal of Asthma. 2013;50(7):776-82.
116. O'Shea MP, Teeling M, Bennett K. An observational study examining the effect of comorbidity on the rates of persistence and adherence to newly initiated oral anti-hyperglycaemic agents. Pharmacoepidemiology and Drug Safety. 2013;22(12):1336-44.
117. Wei YJ, Palumbo FB, Simoni-Wastila L, Shulman LM, Stuart B, Beardsley R, et al. Antiparkinson drug use and adherence in medicare part D beneficiaries with Parkinson's disease. Clinical Therapeutics. 2013;35(10):1513-25.e1.
118. Barber C, Gagnon D, Fonda J, Cho K, Hermos J, Miller M. Assessing the impact of prescribing directives on opioid prescribing practices among Veterans Health Administration providers. Pharmacoepidemiology and Drug Safety. 2017;26(1):40-6.
119. Wu J, Thammakhoune J, Dai W, Koren A, Tcherny-Lessenot S, Wu C, et al. Assessment of dronedarone utilization using US claims databases. Clinical Therapeutics. 2014;36(2):264-72.
120. Robst J. Changes in antipsychotic medication use after implementation of a medicaid mental health carve-out in the US. PharmacoEconomics. 2012;30(5):387-96.
121. Franchi C, Tettamanti M, Pasina L, Djignefa CD, Fortino I, Bortolotti A, et al. Changes in drug prescribing to Italian community-dwelling elderly people: The EPIFARM-Elderly Project 2000-2010. European Journal of Clinical Pharmacology. 2014;70(4):437-43.
122. Baftiu A, Johannessen Landmark C, Rusten IR, Feet SA, Johannessen SI, Larsson PG. Changes in utilisation of antiepileptic drugs in epilepsy and non-epilepsy disorders-a pharmacoepidemiological study and clinical implications. European Journal of Clinical Pharmacology. 2016;72(10):1245-54.
123. Gorst-Rasmussen A, Skjoth F, Larsen TB, Rasmussen LH, Lip GYH, Lane DA. Dabigatran adherence in atrial fibrillation patients during the first year after diagnosis: A nationwide cohort study. Journal of Thrombosis and Haemostasis. 2015;13(4):495-504.
124. Gleason PP, Phillips J, Fenrick BA, Delgado-Riley A, Starner CI. Dalfampridine prior authorization program: A cohort study. Journal of Managed Care Pharmacy. 2013;19(1):18-25.

125. Smolina K, Gladstone E, Morgan SG. Determinants of trends in prescription opioid use in British Columbia, Canada, 2005-2013. Pharmacoepidemiology and Drug Safety. 2016;25(5):553-9.

126. Ahn SH, Choi NK, Kim YJ, Seong JM, Shin JY, Jung SY, et al. Drug persistency of cholinesterase inhibitors for patients with dementia of Alzheimer type in Korea. Archives of Pharmacal Research. 2015;38(6):1255-62.

127. Pottegard A, Poulsen BK, Larsen MD, Hallas J. Dynamics of vitamin K antagonist and new oral anticoagulants use in atrial fibrillation: A Danish drug utilization study. Journal of Thrombosis and Haemostasis. 2014;12(9):1413-8.

128. Huiart L, Bouhnik AD, Rey D, Tarpin C, Cluze C, Bendiane MK, et al. Early discontinuation of tamoxifen intake in younger women with breast cancer: Is it time to rethink the way it is prescribed? European Journal of Cancer. 2012;48(13):1939-46.

129. Hashem MG, Cleary K, Fishman D, Nichols L, Khalid M. Effect of concurrent prescription antiarthralgia pharmacotherapy on persistence to aromatase inhibitors in treatment-naive postmenopausal females. Annals of Pharmacotherapy. 2013;47(1):29-34.

130. De Leon SF, Pauls L, Arya V, Shih SC, Singer J, Wang JJ. Effect of physician participation in a multi-element health information and data exchange program on chronic illness medication adherence. Journal of the American Board of Family Medicine. 2015;28(6):742-9.

131. Manns B, Laupland K, Tonelli M, Gao S, Hemmelgarn B. Evaluating the impact of a novel restricted reimbursement policy for quinolone antibiotics: a time series analysis. BMC health services research. 2012;12(pp 290).

132. Warren JR, Falster MO, Fox D, Jorm L. Factors influencing adherence in long-term use of statins. Pharmacoepidemiology and Drug Safety. 2013;22(12):1298-307.

133. Sjolander M, Eriksson M, Glader EL. Few sex differences in the use of drugs for secondary prevention after stroke: A nationwide observational study. Pharmacoepidemiology and Drug Safety. 2012;21(9):911-9.

134. Lyles CR, Seligman HK, Parker MM, Moffet HH, Adler N, Schillinger D, et al. Financial Strain and Medication Adherence among Diabetes Patients in an Integrated Health Care Delivery System: The Diabetes Study of Northern California (DISTANCE). Health Services Research. 2016;51(2):610-24.

135. Stenman M, Holzmann MJ, Sartipy U. Guideline-directed medical therapy for secondary prevention after coronary artery bypass grafting in patients with depression. IJC Heart and Vessels. 2014;3(pp 37-42).

136. Mahmoudi E, Jensen GA. Has medicare part D reduced racial/ethnic disparities in prescription drug use and spending? Health Services Research. 2014;49(2):502-25.

137. Malo S, Jose Rabanaque M, Feja C, Jesus Lallana M, Aguilar I, Bjerrum L. High antibiotic consumption: a characterization of heavy users in Spain. Basic & Clinical Pharmacology & Toxicology. 2014;115(3):231-6.

138. Schaffer AL, Pearson SA, Buckley NA. How does prescribing for antihypertensive products stack up against guideline recommendations? An Australian population-based study (2006-2014). British Journal of Clinical Pharmacology. 2016;82(4):1134-45.

139. Zarrinkoub R, Kahan T, Johansson SE, Wandell P, Mejhert M, Wettermark B. How to best assess quality of drug treatment in patients with heart failure. European Journal of Clinical Pharmacology. 2016;72(8):965-75.

140. Trotta F, Mayer F, Mecozzi A, Amato L, Addis A. Impact of Guidance on the Prescription Patterns of G-CSFs for the Prevention of Febrile Neutropenia

Following Anticancer Chemotherapy: A Population-Based Utilization Study in the Lazio Region. BioDrugs. 2017;31(2):117-24.

141. Henderson RR, Visaria J, Bridges GG, Dorholt M, Levin RJ, Frazee SG. Impact of specialty pharmacy on telaprevir-containing 3-drug hepatitis C regimen persistence. Journal of Managed Care Pharmacy. 2014;20(12):1227-34.

142. Epstein RA, Bobo WV, Shelton RC, Arbogast PG, Morrow JA, Wang W, et al. Increasing use of atypical antipsychotics and anticonvulsants during pregnancy. Pharmacoepidemiology and Drug Safety. 2013;22(7):794-801.

143. Manteuffel M, Williams S, Chen W, Verbrugge RR, Pittman DG, Steinkellner A. Influence of patient sex and gender on medication use, adherence, and prescribing alignment with guidelines. Journal of Women's Health. 2014;23(2):112-9.

144. Aznar-Lou I, Fernandez A, Gil-Girbau M, Fajo-Pascual M, Moreno-Peral P, Penarrubia-Maria MT, et al. Initial medication non-adherence: prevalence and predictive factors in a cohort of 1.6 million primary care patients. British Journal of Clinical Pharmacology. 2017;83(6):1328-40.

145. Wang LJ, Yang KC, Lee SY, Yang CJ, Huang TS, Lee TL, et al. Initiation and persistence of pharmacotherapy for youths with attention deficit hyperactivity disorder in Taiwan. PLoS ONE. 2016;11(8).

146. Fournier JP, Lapeyre-Mestre M, Sommet A, Dupouy J, Poutrain JC, Montastruc JL. Laboratory monitoring of patients treated with antihypertensive drugs and newly exposed to non steroidal anti-inflammatory drugs: A cohort study. PLoS ONE. 2012;7(3).

147. Huiart L, Ferdynus C, Dell'Aniello S, Bakiri N, Giorgi R, Suissa S. Measuring persistence to hormonal therapy in patients with breast cancer: Accounting for temporary treatment discontinuation. Pharmacoepidemiology and Drug Safety. 2014;23(8):882-9.

148. Berger A, Edelsberg J, Sanders KN, Alvir JMJ, Mychaskiw MA, Oster G. Medication adherence and utilization in patients with schizophrenia or bipolar disorder receiving aripiprazole, quetiapine, or ziprasidone at hospital discharge: A retrospective cohort study. BMC Psychiatry. 2012;12(no pagination).

149. Thorpe CT, Johnson H, Dopp AL, Thorpe JM, Ronk K, Everett CM, et al. Medication oversupply in patients with diabetes. Research In Social & Administrative Pharmacy. 2015;11(3):382-400.

150. Curtis JR, Cai Q, Wade SW, Stolshek BS, Adams JL, Balasubramanian A, et al. Osteoporosis medication adherence: Physician perceptions vs. patients' utilization. Bone. 2013;55(1):1-6.

151. Ferrajolo C, Arcoraci V, Sullo MG, Rafaniello C, Sportiello L, Ferrara R, et al. Pattern of statin use in southern Italian primary care: Can prescription databases be used for monitoring long-term adherence to the treatment? PLoS ONE. 2014;9(7).

152. Bateman BT, Hernandez-Diaz S, Rathmell JP, Seeger JD, Doherty M, Fischer MA, et al. Patterns of opioid utilization in pregnancy in a large cohort of commercial insurance beneficiaries in the United States. Anesthesiology. 2014;120(5):1216-24.

153. Tomas A, Tomic Z, Milijasevic B, Ban M, Horvat O, Vukmirovic S, et al. Patterns of prescription antihypertensive drug utilization and adherence to treatment guidelines in the city of Novi Sad. Vojnosanitetski Pregled. 2016;73(6):531-7.

154. Hackshaw MD, Nagar SP, Parks DC, Miller LA. Persistence and compliance with pazopanib in patients with advanced renal cell carcinoma within a U.S. administrative claims database. Journal of Managed Care & Specialty Pharmacy. 2014;20(6):603-10.

155. Grimmsmann T, Himmel W. Persistence of antihypertensive drug use in German primary care: A follow-up study based on pharmacy claims data. European Journal of Clinical Pharmacology. 2014;70(3):295-301.
156. Kashiwagi K, Furuya T. Persistence with topical glaucoma therapy among newly diagnosed Japanese patients. Japanese Journal of Ophthalmology. 2014;58(1):68-74.
157. Degli Esposti L, Favalli EG, Sangiorgi D, Di Turi R, Farina G, Gambera M, et al. Persistence, switch rates, drug consumption and costs of biological treatment of rheumatoid arthritis: An observational study in Italy. ClinicoEconomics and Outcomes Research. 2017;9(pp 9-17).
158. Handelsman DJ. Pharmacoepidemiology of testosterone prescribing in Australia, 1992-2010. Medical Journal of Australia. 2012;196(10):642-5.
159. Wilk A, Sajjan S, Modi A, Fan CP, Mavros P. Post-fracture pharmacotherapy for women with osteoporotic fracture: analysis of a managed care population in the USA. Osteoporosis International. 2014;25(12):2777-86.
160. Wettermark B, Brandt L, Kieler H, Boden R. Pregabalin is increasingly prescribed for neuropathic pain, generalised anxiety disorder and epilepsy but many patients discontinue treatment. International Journal of Clinical Practice. 2014;68(1):104-10.
161. Genberg BL, Rogers WH, Lee Y, Qato DM, Dore DD, Hutchins DS, et al. Prescriber and pharmacy variation in patient adherence to five medication classes measured using implementation during persistent episodes. Pharmacoepidemiology and Drug Safety. 2016;25(7):790-7.
162. Al Balushi KA, Alzaabi MA, Alghafri F. Prescribing pattern of antifungal medications at a tertiary care hospital in Oman. Journal of Clinical and Diagnostic Research. 2016;10(12).
163. Naldi I, Piccinni C, Mostacci B, Renzini J, Accetta G, Bisulli F, et al. Prescription patterns of antiepileptic drugs in young women: development of a tool to distinguish between epilepsy and psychiatric disorders. Pharmacoepidemiology and Drug Safety. 2016;25(7):763-9.
164. Oymar K, Mikalsen IB, Furu K, Nystad W, Karlstad O. Prescription patterns of inhaled corticosteroids for preschool children--A Norwegian register study. Pediatric Allergy & Immunology. 2016;26(7):655-61.
165. Crijns HJ, van Rein N, Gispen-de Wied CC, Straus SM, de Jong-van den Berg LT. Prescriptive contraceptive use among isotretinoin users in the Netherlands in comparison with non-users: a drug utilisation study. Pharmacoepidemiology & Drug Safety. 2012;21(10):1060-6.
166. Pottegard A, Christensen RD, Houji A, Christiansen CB, Paulsen MS, Thomsen JL, et al. Primary non-adherence in general practice: A Danish register study. European Journal of Clinical Pharmacology. 2014;70(6):757-63.
167. Othman F, Card TR, Crooks CJ. Proton pump inhibitor prescribing patterns in the UK: A primary care database study. Pharmacoepidemiology and Drug Safety. 2016;25:1079-87.
168. Broeks SC, Thisted Horsdal H, Glejsted Ingstrup K, Gasse C. Psychopharmacological drug utilization patterns in pregnant women with bipolar disorder - A nationwide register-based study. Journal of Affective Disorders. 2017;210(pp 158-165).
169. Diene E, Geoffroy-Perez B, Cohidon C, Gauvin S, Carton M, Fouquet A, et al. Psychotropic drug use in a cohort of workers 4 years after an industrial disaster in France. Journal of Traumatic Stress. 2014;27(4):430-7.
170. Haastrup P, Paulsen MS, Zwisler JE, Begtrup LM, Hansen JM, Rasmussen S, et al. Rapidly increasing prescribing of proton pump inhibitors in primary care

despite interventions: A nationwide observational study. European Journal of General Practice. 2014;20(4):290-3.
171. Bergeson JG, Worley K, Louder A, Ward M, Graham J. Retrospective database analysis of the impact of prior authorization for type 2 diabetes medications on health care costs in a medicare advantage prescription drug plan population. Journal of Managed Care Pharmacy. 2013;19(5):374-84.
172. Hsieh KP, Chen LC, Cheung KL, Yang YH. Risks of nonadherence to hormone therapy in Asian women with breast cancer. Kaohsiung Journal of Medical Sciences. 2015;31(6):328-34.
173. Niedrig DF, Gott C, Fischer A, Muller ST, Greil W, Bucklar G, et al. Second-generation antipsychotics in a tertiary care hospital: prescribing patterns, metabolic profiles, and drug interactions. International Clinical Psychopharmacology. 2016;31(1):42-50.
174. Nordin M, Dackehag M, Gerdtham UG. Socioeconomic inequalities in drug utilization for Sweden: Evidence from linked survey and register data. Social Science and Medicine. 2013;77(1):106-17.
175. Martins D, Yao Z, Tadrous M, Shah BR, Juurlink DN, Mamdani MM, et al. The appropriateness and persistence of testosterone replacement therapy in Ontario. Pharmacoepidemiology and Drug Safety. 2017;26(2):119-26.
176. Strom O, Landfeldt E. The association between automatic generic substitution and treatment persistence with oral bisphosphonates. Osteoporosis International. 2012;23(8):2201-8.
177. Baik SH, Rollman BL, Reynolds ICF, Lave JR, Smith KJ, Zhang Y. The effect of the US Medicare Part D coverage gaps on medication use among patients with depression and heart failure. Journal of Mental Health Policy and Economics. 2012;15(3):105-18.
178. Malo S, Bjerrum L, Feja C, Lallana MJ, Abad JM, Rabanaque-Hernandez MJ. The quality of outpatient antimicrobial prescribing: a comparison between two areas of northern and southern Europe. European Journal of Clinical Pharmacology. 2014;70(3):347-53.
179. Pearson SA, Schaffer A. The use and impact of cancer medicines in routine clinical care: methods and observations in a cohort of elderly Australians. BMJ Open. 2014;4(5):e004099.
180. Pottegard A, Bjerregaard BK, Glintborg D, Kortegaard LS, Hallas J, Moreno SI. The use of medication against attention deficit/hyperactivity disorder in Denmark: A drug use study from a patient perspective. European Journal of Clinical Pharmacology. 2013;69(3):589-98.
181. Warle-Van Herwaarden MF, Koffeman AR, Valkhoff VE, Jong GW, Kramers C, Sturkenboom MC, et al. Time-trends in the prescribing of gastroprotective agents to primary care patients initiating low-dose aspirin or non-steroidal anti-inflammatory drugs: a population-based cohort study. British Journal of Clinical Pharmacology. 2015;80(3):589-98.
182. Perrone V, Sangiorgi D, Buda S, Esposti LD. Topical medication utilization and health resources consumption in adult patients affected by psoriasis: Findings from the analysis of administrative databases of local health units. ClinicoEconomics and Outcomes Research. 2017;9(pp 181-188).
183. Pfeiffer PN, Szymanski BR, Valenstein M, McCarthy JF, Zivin K. Trends in antidepressant prescribing for new episodes of depression and implications for health system quality measures. Medical Care. 2012;50(1):86-90.
184. Blume SW, Fox KM, Joseph G, Chuang CC, Thomas J, Gandra SR. Tumor necrosis factor-blocker dose escalation in rheumatoid arthritis patients in a pharmacy benefit management setting. Advances in Therapy. 2013;30(5):517-27.

185. Aguglia E, Ravasio R, Simonetti M, Pecchioli S, Mazzoleni F. Use and treatment modalities for SSRI and SNRI antidepressants in Italy during the period 20032009. Current Medical Research and Opinion. 2012;28(9):1475-84.
186. Haervig KB, Mortensen LH, Hansen AV, Strandberg-Larsen K. Use of ADHD medication during pregnancy from 1999 to 2010: A Danish register-based study. Pharmacoepidemiology and Drug Safety. 2014;23(5):526-33.
187. Bozic B, Bajcetic M. Use of antibiotics in paediatric primary care settings in Serbia. Archives of Disease in Childhood. 2015;100(10):966-9.
188. Schjerning O, Pottegard A, Damkier P, Rosenzweig M, Nielsen J. Use of Pregabalin - A Nationwide Pharmacoepidemiological Drug Utilization Study with Focus on Abuse Potential. Pharmacopsychiatry. 2016;49(4):155-61.
189. Jobski K, Enders D, Amann U, Suzart K, Wallander MA, Schink T, et al. Use of rivaroxaban in Germany: A database drug utilization study of a drug started in hospital. European Journal of Clinical Pharmacology. 2014;70(8):975-81.
190. Fitch K, Broulette J, Pyenson BS, Iwasaki K, Kwong WJ. Utilization of anticoagulation therapy in medicare patients with nonvalvular atrial fibrillation. American Health and Drug Benefits. 2012;5(3):157-68.
191. Peng X, Wu N, Chen SY, Yu X, Andrews JS, Novick D. Utilization of duloxetine and celecoxib in osteoarthritis patients. Current Medical Research and Opinion. 2013;29(9):1161-9.
192. Chang CM, Wu CS, Huang YW, Chau YL, Tsai HJ. Utilization of psychopharmacological treatment among patients with newly diagnosed bipolar disorder from 2001 to 2010. Journal of Clinical Psychopharmacology. 2016;36(1):32-44.
193. Arfe A, Nicotra F, Ghirardi A, Simonetti M, Lapi F, Sturkenboom M, et al. A probabilistic bias analysis for misclassified categorical exposures, with application to oral anti-hyperglycaemic drugs. Pharmacoepidemiology and Drug Safety. 2016;25(12):1443-50.
194. Grimaldi-Bensouda L, Rossignol M, Aubrun E, Benichou J, Abenhaim L. Agreement between patients' self-report and physicians' prescriptions on nonsteroidal anti-inflammatory drugs and other drugs used in musculoskeletal disorders: The international Pharmacoepidemiologic General Research eXtension database. Pharmacoepidemiology and Drug Safety. 2012;21(7):753-9.
195. Pottegard A, Hallas J. Assigning exposure duration to single prescriptions by use of the waiting time distribution. Pharmacoepidemiology & Drug Safety. 2013;22(8):803-9.
196. De Groot MCH, Candore G, Uddin MJ, Souverein PC, Ali MS, Belitser SV, et al. Case-only designs for studying the association of antidepressants and hip or femur fracture. Pharmacoepidemiology and Drug Safety. 2016;25(pp 103-113).
197. Miller TP, Troxel AB, Li Y, Huang YS, Alonzo TA, Gerbing RB, et al. Comparison of administrative/billing data to expected protocol-mandated chemotherapy exposure in children with acute myeloid leukemia: A report from the Children's Oncology Group. Pediatric Blood and Cancer. 2015;62(7):1184-9.
198. Lauffenburger JC, Balasubramanian A, Farley JF, Critchlow CW, O'Malley CD, Roth MT, et al. Completeness of prescription information in US commercial claims databases. Pharmacoepidemiology and Drug Safety. 2013;22(8):899-906.
199. Wallach Kildemoes H, Hendriksen C, Andersen M. Drug utilization according to reason for prescribing: A pharmacoepidemiologic method based on an indication hierarchy. Pharmacoepidemiology and Drug Safety. 2012;21(10):1027-35.
200. Bijlsma MJ, Janssen F, Hak E. Estimating time-varying drug adherence using electronic records: extending the proportion of days covered (PDC) method. Pharmacoepidemiology & Drug Safety. 2016;25(3):325-32.

201. Gamble JM, Johnson JA, Majumdar SR, McAlister FA, Simpson SH, Eurich DT. Evaluating the introduction of a computerized prior-authorization system on the completeness of drug exposure data. Pharmacoepidemiology and Drug Safety. 2013;22(5):551-5.

202. Lum KJ, Newcomb CW, Roy JA, Carbonari DM, Saine ME, Cardillo S, et al. Evaluation of methods to estimate missing days' supply within pharmacy data of the Clinical Practice Research Datalink (CPRD) and The Health Improvement Network (THIN). European Journal of Clinical Pharmacology. 2017;73(1):115-23.

203. Tanskanen A, Taipale H, Koponen M, Tolppanen AM, Hartikainen S, Ahonen R, et al. From prescription drug purchases to drug use periods - a second generation method (PRE2DUP). BMC Med Inform Decis Mak. 2015;15:21.

204. Yun H, Curtis JR, Saag K, Kilgore M, Muntner P, Smith W, et al. Generic alendronate use among medicare beneficiaries: Are part d data complete? Pharmacoepidemiology and Drug Safety. 2013;22(1):55-63.

205. Hurault-Delarue C, Chouquet C, Savy N, Lacroix I, Beau AB, Montastruc JL, et al. How to take into account exposure to drugs over time in pharmacoepidemiology studies of pregnant women? Pharmacoepidemiology and Drug Safety. 2016;25(7):770-7.

206. Nielen JT, van den Bemt BJ, Boonen A, Dagnelie PC, Emans PJ, Veldhorst N, et al. Identification of antithrombotic drugs related to total joint replacement using anonymised free-text notes: a search strategy in the Clinical Practice Research Datalink. BMJ Open. 2015;5(11):e009017.

207. Gamble JM, Johnson JA, McAlister FA, Majumdar SR, Simpson SH, Eurich DT. Limited Impact of Drug Exposure Misclassification From Non-Benefit Thiazolidinedione Drug Use on Mortality and Hospitalizations From Saskatchewan, Canada: A Cohort Study. Clinical Therapeutics. 2015;37(3):629-42.

208. Burne RM, Abrahamowicz M. Martingale residual-based method to control for confounders measured only in a validation sample in time-to-event analysis. Statistics in Medicine. 2016;35(25):4588-606.

209. Sinnott SJ, Polinski JM, Byrne S, Gagne JJ. Measuring drug exposure: Concordance between defined daily dose and days' supply depended on drug class. Journal of Clinical Epidemiology. 2016;69(pp 107-113).

210. Sauer B, Nebeker J, Shen S, Rupper R, West S, Shinogle JA, et al. Methodological framework to identify possible adverse drug reactions using population-based administrative data. F1000Research. 2014;3:258.

211. Van Le H, Beach KJ, Powell G, Pattishall E, Ryan P, Mera RM. Performance of a semi-automated approach for risk estimation using a common data model for longitudinal healthcare databases. Statistical Methods in Medical Research. 2013;22(1):97-112.

212. Tamblyn R, Girard N, Dixon WG, Haas J, Bates DW, Sheppard T, et al. Pharmacosurveillance without borders: electronic health records in different countries can be used to address important methodological issues in estimating the risk of adverse events. Journal of Clinical Epidemiology. 2016;77(pp 101-111).

213. Leppee M, Boskovic J, Culig J, Eric M. Pharmacy claims data as a tool to measure adherence. Current Medical Research and Opinion. 2012;28(8):1389-93.

214. Gamble JM, McAlister FA, Johnson JA, Eurich DT. Quantifying the impact of drug exposure misclassification due to restrictive drug coverage in administrative databases: A simulation cohort study. Value in Health. 2012;15(1):191-7.

215. Laforest L, Licaj I, Devouassoux G, Chatte G, Belhassen M, Van Ganse E, et al. Relative exposure to controller therapy and asthma exacerbations: A

validation study in community pharmacies. Pharmacoepidemiology and Drug Safety. 2014;23(9):958-64.
216. Gamble JM, McAlister FA, Johnson JA, Eurich DT. Restrictive drug coverage policies can induce substantial drug exposure misclassification in pharmacoepidemiologic studies. Clinical Therapeutics. 2012;34(6):1379-86.e3.
217. Burden AM, Huang A, Tadrous M, Cadarette SM. Variation in the days supply field for osteoporosis medications in Ontario. Archives of Osteoporosis. 2013;8(1-2).
218. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. Journal of the American Medical Informatics Association. 2010;17(6):652-62.
219. Pazzagli L, Brandt L, Linder M, Myers D, Mavros P, Andersen M, et al. Methods for constructing treatment episodes and impact on exposure-outcome associations. European Journal of Clinical Pharmacology. 2019;76 pag:267 – 75.
220. Eskin M, Simpson SH, Eurich DT. Impact of drug exposure definitions on observed associations in pharmacoepidemiology research. Journal of Population Therapeutics and Clinical Pharmacology. 2018;25(1):E39-E52.
221. Langan SM, Schmidt SAJ, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). Bmj-British Medical Journal. 2018;363.
222. Wallerstedt SM, Wettermark B, Hoffmann M. The First Decade with the Swedish Prescribed Drug Register - A Systematic Review of the Output in the Scientific Literature. Basic & Clinical Pharmacology & Toxicology. 2016;119(5):464-9.
223. Richter H, Dombrowski S, Hamer H, Hadji P, Kostev K. Use of a German longitudinal prescription database (LRx) in pharmacoepidemiology. German medical science : GMS e-journal. 2015;13:Doc14-Doc.
224. Tseng C-H. A Review on Thiazolidinediones and Bladder Cancer in Human Studies. Journal of Environmental Science and Health Part C-Environmental Carcinogenesis & Ecotoxicology Reviews. 2014;32(1):1-45.
225. Ramsey RR, Ryan JL, Hershey AD, Powers SW, Aylward BS, Hommel KA. Treatment Adherence in Patients With Headache: A Systematic Review. Headache. 2014;54(5):795-816.
226. Taipale H, Tanskanen A, Koponen M, Tolppanen AM, Tiihonen J, Hartikainen S. Agreement between PRE2DUP register data modeling method and comprehensive drug use interview among older persons. Clin Epidemiol. 2016;8:363-71.
227. Gagne JJ. Restrictive Reimbursement Policies: Bias Implications for Claims-Based Drug Safety Studies. Drug Safety. 2014;37(10):771-6.
228. Hampp C, Greene P, Pinheiro SP. Use of Prescription Drug Samples in the USA: A Descriptive Study with Considerations for Pharmacoepidemiology. Drug Safety. 2016;39(3):261-70.
229. Lohr S. For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. 2014 20140817.
230. Williams R, Brown B, Peek N, Buchan I. Making Medication Data Meaningful: Illustrated with Hypertension. Stud Health Technol Inform. 2016;228:247-51.
231. Shah AD, Martinez C. An algorithm to derive a numerical daily dose from unstructured text dosage instructions. Pharmacoepidemiology and Drug Safety. 2006;15(3):161-6.

232. Menditto E, De Gea AB, Cahir C, Marengoni A, Riegler S, Fico G, et al. Scaling up health knowledge at European level requires sharing integrated data: An approach for collection of database specification. ClinicoEconomics and Outcomes Research. 2016;8(pp 253-265).

233. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International Journal of Epidemiology. 2015;44(3):827-36.

234. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, et al. Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. Pharmacoepidemiology and Drug Safety. 2017;26(9):1018-32.

235. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. 1996;5(3):299-314.

236. Venables WN, Smith DM, Team RC. An Introduction to R 2018 [Available from: https://cran.r-project.org/manuals.html, last accessed 13/06/2020.

237. Peng RD. R Programming for Data Science 2019. Available from: https://bookdown.org/rdpeng/rprogdatascience/, last accessed 13/06/2020.

238. Wickham H. Advanced R. Boca Raton, FL: CRC Press; 2015. Available from: http://adv-r.had.co.nz/, last accessed 13/06/2020.

239. RStudio. RStudio IDE Features 2020 [Available from: https://rstudio.com/products/rstudio/features/, last accessed 13/06/2020.

240. Comprehensive R Archive Network. CRAN - Contributed Packages 2020 [Available from: https://cran.r-project.org/web/packages/index.html, last accessed 13/06/2020.

241. Dima A. Computation of adherence to medication and visualization of medication histories in R with AdhereR: Towards transparent and reproducible use of electronic healthcare data. PLoS One. 2018;12(4):e0174426.

242. ISD Scotland. Prescribing and Medicines - PIS Fields for Researchers 2020 [v5:[Available from: https://www.isdscotland.org/Health-Topics/Prescribing-and-Medicines/Prescribing-Datamarts/docs/PIS_fields_for_researchers_v5_eDRIS%20Guidance.pdf, last accessed 13/06/2020.

243. WHO Collaborating Centre for Drug Statistics Methodology. WHOCC - ATC/DDD Index 2019 [Available from: https://www.whocc.no/atc_ddd_index/, last accessed 13/06/2020.

244. World Health Organisation. Epilepsy - Fact sheet 2019 [Available from: https://www.who.int/news-room/fact-sheets/detail/epilepsy, last accessed 13/06/2020.

245. National Institute for Health and Care Excellence. Epilepsies: diagnosis and management: NICE; 2012 [Available from: https://www.nice.org.uk/guidance/cg137, last accessed 13/06/2020.

246. Chang BS, Lowenstein DH. Mechanisms of disease - Epilepsy. New England Journal of Medicine. 2003;349(13):1257-66.

247. Ridsdale L, Charlton J, Ashworth M, Richardson MP, Gulliford MC. Epilepsy mortality and risk factors for death in epilepsy: a population-based study. British Journal of General Practice. 2011;61(586).

248. Weatherburn CJ, Heath CA, Mercer SW, Guthrie B. Physical and mental health comorbidities of epilepsy: Population-based cross-sectional analysis of 1.5 million people in Scotland. Seizure-European Journal of Epilepsy. 2017;45:125-31.

249. Tellez-Zenteno JF, Patten SB, Jette N, Williams J, Wiebe S. Psychiatric comorbidity in epilepsy: A population-based analysis. Epilepsia. 2007;48(12):2336-44.

250. Olesen JB, Abildstrom SZ, Erdal J, Gislason GH, Weeke P, Andersson C, et al. Effects of epilepsy and selected antiepileptic drugs on risk of myocardial infarction, stroke, and death in patients with or without previous stroke: a nationwide cohort study. Pharmacoepidemiology and Drug Safety. 2011;20(9):964-71.
251. Stefan H, Feuerstein TJ. Novel anticonvulsant drugs. Pharmacology & Therapeutics. 2007;113(1):165-83.
252. LaRoche SM, Helmers SL. The new antiepileptic drugs - Clinical applications. Jama-Journal of the American Medical Association. 2004;291(5):615-20.
253. Krakowski MD. Antiepileptic Drugs - Therapeutic Drug Monitoring of the Newer Generation Drugs: Clinical Laboratory News; 2013 [Available from: https://www.aacc.org/publications/cln/articles/2013/june/antiepileptic-drugs, last accessed 13/06/2020.
254. Nicholas JM, Ridsdale L, Richardson MP, Ashworth M, Gulliford MC. Trends in antiepileptic drug utilisation in UK primary care 1993-2008: Cohort study using the General Practice Research Database. Seizure-European Journal of Epilepsy. 2012;21(6):466-70.
255. Powell G, Logan J, Kiri V, Borghs S. Trends in antiepileptic drug treatment and effectiveness in clinical practice in England from 2003 to 2016: a retrospective cohort study using electronic medical records. Bmj Open. 2019;9(12).
256. Bolin K, Berggren F, Berling P, Morberg S, Gauffin H, Landtblom AM. Patterns of antiepileptic drug prescription in Sweden: A register-based approach. Acta Neurologica Scandinavica. 2017;136(5):521-7.
257. Hollingworth SA, Eadie MJ. Antiepileptic drugs in Australia: 2002-2007. Pharmacoepidemiology and Drug Safety. 2010;19(1):82-9.
258. Perucca E, Tomson T. The pharmacological treatment of epilepsy in adults. Lancet Neurology. 2011;10(5):446-56.
259. Lai EC-C, Hsieh C-Y, Su C-C, Yang Y-HK, Huang C-W, Lin S-J, et al. Comparative persistence of antiepileptic drugs in patients with epilepsy: A STROBE-compliant retrospective cohort study. Medicine. 2016;95(35).
260. Jacob L, Hamer HM, Kostev K. Persistence with antiepileptic drugs in epilepsy patients treated in neurological practices in Germany. Epilepsy & Behavior. 2017;73:204-7.
261. NHS Research Scotland. Safe Havens | NHS Research Scotland 2020 [Available from: https://www.nhsresearchscotland.org.uk/research-in-scotland/data/safe-havens, last accessed 13/06/2020.
262. NHS Greater Glasgow and Clyde Communications. Available Data in the NHS Greater Glasgow & Clyde Safe Haven 2020 [Available from: https://www.nhsggc.org.uk/media/260115/glasgow-safe-haven-dataset-catalogue.pdf, last accessed 13/06/2020.
263. World Health Organisation. ICD-10 : international statistical classification of diseases and related health problems, tenth revision: World Health Organization; 2016. Available from: https://apps.who.int/iris/handle/10665/42980, last accessed 13/06/2020.
264. Gasparini A. comorbidity: An R package for computing comorbidity scores. Journal of Open Source Software. 2020;3(23):648.
265. Quan HD, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical Care. 2005;43(11):1130-9.
266. Mbizvo GK, Dixon P, Hutton JL, Marson AG. Levetiracetam add-on for drug-resistant focal epilepsy: an updated Cochrane Review. Cochrane Database Syst Rev. 2012;2012(9).

267. GlaxoSmithKline. Policies, codes and standards - Evergreening: GSK; 2019 [Available from: https://www.gsk.com/media/2949/evergreening-policy.pdf, last accessed 13/06/2020.

268. Agency MaHpR. Valproate use by women and girls: United Kingdom; 2018 [updated 24/01/2020. Available from: https://www.gov.uk/guidance/valproate-use-by-women-and-girls, last accessed 13/06/2020.

269. Pazzagli L, Linder M, Zhang M, Vago E, Stang P, Myers D, et al. Methods for time-varying exposure related problems in pharmacoepidemiology: An overview. Pharmacoepidemiology and Drug Safety. 2018;27(2):148-60.

270. Faught E, Duh MS, Weiner JR, Guerin A, Cunnington MC. Nonadherence to antiepileptic drugs and increased mortality Findings from the RANSOM Study. Neurology. 2008;71(20):1572-8.

271. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. International Journal of Surgery. 2014;12(12):1495-9.

272. Pye SR, Sheppard T, Joseph RM, Lunt M, Girard N, Haas JS, et al. Assumptions made when preparing drug exposure data for analysis have an impact on results: An unreported step in pharmacoepidemiology studies. Pharmacoepidemiology and Drug Safety. 2018;27(7):781-8.