Death recording in the Clinical Practice Research Datalink

Arlene Gallagher

Death recording in the Clinical Practice Research Datalink

Arlene Gallagher

This research uses data provided by patients and collected by the National Health Service in the UK as part of their care and support.

Printed by: Proefschriftmaken.nl

ISBN: 978 94 6380 533 9

Death recording in the Clinical Practice Research Datalink

Registratie van overlijden in de 'Clinical Practice Research Datalink'
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op 21 Oktober 2019 des middags te 12.45 uur

door

Arlene Marie Gallagher

geboren op 10 augustus 1979
te London, Verenigd Koninkrijk

Promotoren: Prof. dr. H.G.M. Leufkens
Prof. dr. F. de Vries

The research presented in this PhD thesis was conducted under the umbrella of the Utrecht-World Health Organization (WHO) Collaborating Centre for Pharmaceutical Policy and Regulation (http://www.pharmaceuticalpolicy.nl/), located at the Division of Pharmacoepidemiology and Clinical Pharmacology of Utrecht University in the Netherlands and with financial support from the Clinical Practice Research Datalink.

This thesis is dedicated to the memory of
Dannie Gallagher
and
Vincent Joseph Stakelum

Table of contents

Chapter 1:

General Introduction

Pharmacoepidemiology using primary care electronic healthcare records: death data

The main aim of research in pharmacoepidemiology is to assess the safety, efficacy or effectiveness of treatments. The two basic components of any study are exposure and outcome; the exposure is often a treatment or intervention, and the outcome is usually cure, symptom relief or survival [1]. The outcome, or endpoint, of a study documents the impact of an exposure or intervention on the health of a given population. The primary outcome is the variable that is the most relevant to answer the research question. Secondary outcomes lend supporting evidence to the primary outcome. Death is seen as a hard outcome; it is well-defined in clinical practice, with limited ambiguity. However, analysing mortality as a primary or secondary outcome in research relies upon accurate and reliable recording of both the vital status (dead versus alive) and the associated date.

The founding fathers of academic primary care conducted research using paper-based routinely collected practice data [2]. With the development of digital data collection and innovation in information technology, there has been a surge in the use of electronic healthcare records (EHR) for research over the past 30 years [3]. These data sources draw on the routine collection of data from everyday clinical care and enable researchers to access information for a large number of patients, at relatively high speed and low cost [4]. However, some EHRs are based in the primary care setting and others from secondary care hospitals; some include the national population and others are based on a sample, which may or may not be representative of the whole country. The quality of the data vary considerably and methodological developments have been necessary to deal with missing data and various forms of bias in such databases [5].

Primary care EHR data have often been used for analyses with a death outcome [6-13], however information on vital status and the date of death may be frequently incomplete since the data are recorded for the purpose of ongoing clinical care. A General Practitioner (GP) is not required to report the fact that death has occurred under English law [14] and the completion of the death certificate could be authorised by professionals other than the GP. The circumstances of death may have an impact on the likelihood of records being updated within good time or being missing altogether, for example when an individual dies outside of hospital and is taken straight to funeral director or when information on the causes and events leading to the death do not reach the GP before the patient record is closed. In the UK, as many as 50% of deaths occur in hospital [15, 16], however hospitalised events are not generally captured in the primary care record unless discharge details are subsequently added. The recording of vital status and the date of death at the GP practice relies upon a relay of information from hospitals, coroners and family

members and previous research has shown there is delay in informing GPs of patient deaths [17]. Whilst information on the cause of death has been reported as inaccurate or incomplete [18, 19], at the time of starting this thesis there had not been any studies that assessed the accuracy of the status of death and the recorded date in primary care electronic healthcare records.

Death as a point of right-censoring in observational studies

Records on vital status are crucial to pharmacoepidemiology not just for studies with a mortality outcome but also as a censoring end point for all patients in a study. The ability to accurately determine follow-up time is pivotal in epidemiological evaluations. Missing information on a key right-censoring variable such as death is likely to have an impact on the robustness of a study as this could result in the erroneous inclusion or exclusion of individuals in the study population and/or contribute to a miscalculation of person time. The inclusion of follow-up time after an individual has died inflates the denominator for a study and this type of information bias cannot be corrected for with analytic techniques.

Death as an outcome in pharmacoepidemiology

Studying mortality in pharmacoepidemiology can be challenging; researchers need to consider the size of the study population and the healthcare setting. First, death is usually a rare outcome in pharmacoepidemiology. Large study populations may be required to meet sample size estimations and therefore mortality may be pooled with other outcomes, or treated as a secondary outcome, if the power of the study to detect a difference in mortality is low. Second, the choice of study setting may have an impact on the generalisability of the results. Analysing deaths in a database of hospital records is likely to include a more severe population than a cohort from primary care. Crane *et al.* (1992) [20] studied deaths in a hospital-based population of asthma patients. The population were all survivors of previous hospitalisations therefore the association between the markers of asthma severity and asthma deaths overall was not generalisable to all asthma patients. In another hospital-based study, researchers analysed the risk of mortality associated with haloperidol compared with atypical antipsychotics in patients admitted to hospital [21]. Patients were followed for deaths over seven days following treatment initiation. The large cohort provided sufficient statistical power to detect a modest difference in mortality however, the study population were all hospital-based and were likely to be different from the population considered at risk of mortality with haloperidol use - elderly people with dementia. EHR databases can offer large populations in various settings. Researchers need to consider the appropriateness of the data source in terms of the likely number of patients and the generalisability of the healthcare setting when designing a study with death as an outcome.

The use of Electronic Health Records (EHR) for pharmacoepidemiology

Large EHR databases are available internationally, varying in size, the type of data collected, the detail of information they contain and the representativeness of the population in question. Key factors include the healthcare system in place for the population, the method of recording data, i.e. coding versus free text and the ability to link with other data sources.

EHR data sources across Europe

In many European countries there is universal healthcare, a system where the GP is the gatekeeper for specialist care, and patients have equal access to largely all care and drug prescribing regardless of income or socioeconomic status. The definition of primary care differs across countries and the role of family medicine varies, however in the main, general practice is the first place for individuals to present their health problems. This type of healthcare system avoids the fragmented arrangement found in other countries [22] and lends itself well to the development of a research database. Systems where inhabitants are registered at, and attend, only one general practice at any one time additionally offer the opportunity to identify control patients for case–control studies and unexposed individuals for cohort studies from the same practice as the cases, which lessens the potential for bias due to practice variation. Examples of healthcare systems like this include the Netherlands, Denmark, Sweden and the UK, which are described next.

In the Netherlands, all inhabitants are listed with a general practice and generally the GP is the first professional to be consulted for health problems [23]. GPs are expected to record diagnostic information using the International Classification of Primary Care version 1 (ICPC-1) but can additionally add free text. There are primary care databases with considerable sample sizes, such as the Integrated Primary Care Information (IPCI) database and the database of the Netherlands Information Network of General Practice (Landelijk Informatienetwerk Huisartsen (LINH), www.linh.nl). The IPCI contains electronic medical records from a representative sample (n=750) of GPs in the Netherlands [24]. The IPCI project (www.ipci.nl) was started in 1989 and contains up to 10 years of observational data from over 2 million individuals. Data are collected from 9 different GP systems that are standardised into one common data structure with many quality control steps before the data are considered research ready. IPCI data have been used to study antimicrobial drugs [25], asthma in children [26] and priapism [27] among others. LINH includes routinely extracted data from a representative sample of 90 general practices using 5 different GP systems, with approximately 350,000 listed patients since 2001 [28]. The LINH database has been for various studies including patients with inflammatory arthritis [29] and antihypertensive drug users [30]. In addition to primary care databases,

there are prescription databases, some of which have links to other healthcare data sources. Information on prescribed medication from public pharmacies in Northern and Eastern Netherlands for a population of approximately 700,000 persons have been collated since 1999 into the InterAction Database IADB.nl [31]. This data source has been used to study treatments during pregnancy [32] and drug utilisation studies of antibiotics [33, 34] among others. The PHARMO Database Network [Institute for Drug Outcome Research, www.pharmo.nl] includes several electronic healthcare databases covering primary care, outpatient pharmacy, hospital data and death registration data (without cause of death), linked at the patient level. The PHARMO record linkage system uses a combination of deterministic (e.g., GP data), semideterministic (e.g., clinical laboratory data), and probabilistic (e.g., hospital data) techniques to link databases to a central patient roster file, with population coverage of 3.2 million individuals [35]. The PHARMO record linkage system has been used in a wide range of studies including myocardial infarction [36] and cancer [37].

All Danish residents are entitled to publicly funded healthcare, which is predominantly free of charge at the point of use. Individuals are required to register with a GP who provides primary care as well as playing a gatekeeping role for access to hospital and specialist care. GPs often resume the responsibility for treatment after a diagnosis has been established by a specialist, therefore GPs issue most prescriptions in Denmark. The National Prescription Registry (NPR) contains individual-level data on all prescriptions filled at Danish community pharmacies since 1995 [38]. A centralised civil registration system has been in place for many years, allowing for the possibility of linkage between all national registries containing civil registration numbers, e.g., patient registry, cancer registry, prescription databases, and registry of causes of death, however there is limited information on treatment indication. Data from the NPR are available to researchers through Statistics Denmark and the Danish Health Data Authority. NPR data have been used extensively in pharmacoepidemiological research alongside linked data sources [39].

In Sweden, patients under the age of 18, as well as some other targeted groups, are exempt from charges for visits to the GP, although others are expected to pay part of the cost for GP services at the point of care. Patient records are kept electronically and ePrescriptions are used with few exceptions. The Swedish Prescribed Drug Register (SPDR) contains information about age, sex and unique identifier of the patient as well as the prescriber's profession and practice [40]. Like Denmark, a centralised civil registration system allows linkage between data sources and the coverage is the entire population, however there is limited information on diagnosis or severity of the conditions treated as the indication for the prescription is generally not recorded [41]. The SPDR have been used mostly for studies of psychiatric and cardiovascular diseases [42].

The National Health Service (NHS) in the UK is publicly funded mainly through general taxation, and a smaller proportion from national insurance, taken through the payroll system. Some of the population also buy supplementary cover for more rapid and convenient access to services, although this is unlikely to be comprehensive as private insurance does not generally cover general practice, maternity care or Accident and Emergency. Over 98% of the UK population are registered with a GP and patients are not expected to pay any upfront costs for visits. Anyone, regardless of nationality and residential status may register and consult with a GP without charge [43] and the GP is the gatekeeper to specialist care. In England, there are some charges for prescriptions, whereas these are currently free of charge in Scotland, Wales and Northern Ireland. Data are collected across a variety of software systems, the current options are Egton Medical Information Systems (EMIS), used by 56% of GP practices in 2016, TPP SystmOne (34%) and Vision (9%) [44]. Multiple data repositories have been developed for research, some of which were based on software that is no longer available, such as the Doctors Independent Network (DIN) and the IMS Disease Analyzer-Mediplus databases, which were based on the iSoft-Torex clinical system [45]. Current available databases for research are as follows:

- QResearch was established in 2004 as a partnership between University of Nottingham and EMIS Health (https://www.emishealth.com) [46]. Data are collected as part of routine care from approximately 1500 general practices using the EMIS clinical computer system. QResearch is now based at Oxford University. The database is open to academic researchers with ethical committee approval. The database has been linked to hospital admission records, cancer registries and death certificates, however these data can only be accessed for analysis on site at the University of Oxford.
- The Clinical Practice Research Datalink (CPRD) Aurum database is also based on practices using EMIS software, launched in October 2017 [47]. Primary care data in CPRD Aurum have been linked to national secondary care databases as well as deprivation and national death registration data [48].
- The ResearchOne database (http://www.researchone.org) consists of de-identified clinical and administrative data drawn from the electronic patient records currently held on the TPP SystmOne clinical system. The database was launched in 2012 as part of a collaboration between TPP and the University of Leeds. The healthcare research data are from general practice, community care, child health, palliative hospital, acute hospital, urgent care, accident & emergency and out-of-hours, however there is no linkage to death certificates [49].
- A new database is currently under development by CPRD, also based on practices using TPP SystmOne software, which will include linkage with other data sources however this is not yet available for research [50].

- The Health Improvement Network (THIN) is based on data collected from practices using the Vision software system. THIN data collection commenced in 2003 as part of a collaborative project between the owners of the software (In Practice Systems Ltd (INPS)) and a global technology and services company (Cegedim, which was later acquired by IMS Health (now IQVIA)). THIN data are representative of the UK in terms of demographics, major condition prevalence and death rates [51, 52].
- The CPRD GOLD database is also based on practices using the Vision software system and will be the primary care data source of interest for this thesis.

CPRD primary care data

A database of records from UK primary care was set up in 1987 and launched as a powerful new information resource for pharmacoepidemiology and pharmacovigilance [53-55]. Originally known as the Value Added Medical Products (VAMP) research databank, free hardware and software were supplied in return for standardised data recording and provision of patient-level data into a central database to be used for health research. Practices were supplied with recording guidelines and received data quality advice and feedback reports based on analysis of the received data [56]. Diagnoses and symptoms were initially recorded using OXMIS codes until newer versions of the practice software system (Vision) were made available and the Read coding system was introduced [57]. Version 3 of the Read codes was designed to cope with the increased detail required by clinical specialists [58], and although at least one Read code is required to record clinical information to the EHR, free text can be additionally included to supplement the coding. The database was renamed the General Practice Research Database (GPRD) in 1993 when VAMP was bought out by Reuters Health Information, and was eventually passed to the Department of Health where it is currently housed by the Medicines and Healthcare Products Regulatory Agency. In 2012 the database was expanded to a national linkage service and renamed the Clinical Practice Research Datalink (CPRD). The original database of primary care records became known as CPRD GOLD.

The data source has an international reputation in the field of drug safety [59] and has been used extensively for pharmacoepidemiology and clinical epidemiology research. Over time, value added information has been supplemented such as data quality measures and hierarchical dictionaries to support data retrieval [60].

Researchers from the UK, Netherlands and elsewhere commonly use this data source for pharmacoepidemiology partly due to the richness of coding and the availability of laboratory test data. In addition, the introduction of the Quality and

Outcomes Framework (QOF) in 2004, a voluntary annual reward and incentive programme for improving chronic disease management in the UK [61], has meant that patients in this database have increased recording on lifestyle factors such as smoking status and Body Mass Index. Analyses of these data are regularly requested by international medicines regulators for periodic safety update reports and post-authorisation safety studies.

The CPRD GOLD database has also been used for high-impact research, such as:

- Demonstrating the absence of an association between measles, mumps and rubella (MMR) vaccine and autism [62]
- Informing NICE (the National Institute for Health and Care Excellence) guidelines by identifying predictors of bladder cancer in general practice [63]
- Linking body mass index (BMI) and risk of cancer [64]

Previous research from Utrecht University has included:

- Analysing the pattern of risk of cancer in patients with metal-on-metal hip replacements [65]
- Assessing the risk of fracture in type 2 diabetes mellitus patients [66] and those using beta blockers [67]

Large collections of electronic medical records from primary care, such as CPRD GOLD are enormously valuable, offering rapid access to data collected in a standardised format [68]. The large number of individuals included in CPRD, together with the longitudinal follow-up time, enable research on rare outcomes and long latency periods [69, 70]. However, the primary purpose of the capture of electronic healthcare records at the GP practice is to support the clinical care of the patient and not for the purposes of research; the recording of information outside of this scope may be missing or only partially recorded and external validation has been recommended [53, 71, 72]. The validity of the coded diagnostic data has been assessed in two systematic reviews, both finding the overall validity to be high [73, 74], however, difficulties have been highlighted with studying events in an outpatient setting and with examining sudden death [75]. There have been numerous updates to the database over the years however at the time of starting this PhD an up to date profile of the CPRD GOLD and its capabilities had not been published.

Death recording in UK primary care

There is no direct linkage between the GP practice and the national registry of deaths, and since the CPRD database is based on records originally recorded for the purposes of clinical care, there may be limited incentive for correct recording of vital status and the date of death. The accuracy of the death date relies on correct and

careful documentation, which is not always a top priority for a busy clinician [76]. Due to the gatekeeper system in the UK, a GP should be informed about all medical events, and deaths. Although general practitioners are involved in the care of most dying patients, they do not routinely receive information about their deceased patients for whom they did not complete the death certificate, and often they rely upon informal communication channels [77]. There may be delays in receiving information about deaths and it is not clear whether all deaths are recorded in the EHR with the official date. Given there is the potential for a death to be missing from the record, erroneously recorded with the current date, or recorded with the wrong date due to misinformation or a typographical error, the vital status and date of death information in the primary care record is subject to question.

In the UK, a GP would usually close the medical record for a patient when notified of their death. At this point, they may or may not choose to record information from the death certificate into the patient's electronic health record. The GPRD recording guidelines stated: "When a patient dies, the DATE of death, the FACT of death, and the CAUSE(S) of death must be recorded correctly, when known. These are crucial pieces of information for many types of research" [56]. This emphasized to GPs the importance of vital status and the date of death for research studies. However, in the Vision system, there is more than one way that death information can be recorded [78]. A clinical record relating to death may be added to the patient's medical record using a Read code. A structured data area, specifically for entering death details (Death Administration screen) may be used to record the date of death and the cause. Alternatively, the patient's electronic health record could be closed, where the registration status of the patient is changed to 'Transferred out' and the reason is recorded as 'Death', with an associated date. Researchers would therefore need to examine four sets of records in the CPRD GOLD database in order to identify potential options for the date of death:

1. The date associated with a Read code related to death, e.g. 8HG..00: "Died in hospital".
2. The date of death as entered in the Death Administration screen.
3. The date that a record was entered in the Death Administration screen.
4. The date the registration status was changed to 'Transferred out', where the reason was 'Death'.

In order to support researchers with this task, an algorithm was developed to identify the best estimate for the date of death based on these four sets of records. The resulting date was added to the patient record as the 'deathdate' field (CPRD death date).

A PubMed search for studies published in 2010 using the terms 'GPRD' and 'death' or 'mortality' resulted in 21 papers. These were predominantly cohort studies, and six mentioned using the available information on death to right-censor their study population [79-84]; this was assumed to be the CPRD death date. Most of the papers did not describe the records that were used to identify death or outline the reliability of the death information in the limitations. Eight studies examined death as an outcome [6-13], only one of which provided some details on how death was identified, using a combination of three methods, similar to the algorithm used by CPRD [11]. Two of the articles stated that the mortality data in GPRD has been validated, referencing Shah and Martinez (2004) [85]. However, this reference is to a conference abstract with limited details; there is no full paper describing the validation.

At the time of starting this PhD in 2011, the algorithm had not been validated, the methodology had not been published, and the accuracy of the vital status and recording of the date of death for patients in CPRD GOLD had not been examined.

Validating UK primary care death data

Validation of diagnoses is generally recommended if there is any doubt about the specificity of the coding of the condition [4], and this could also apply to death information. However, death information is rarely subject to individual verification unless a study is investigating the specific cause of death or analysing sudden deaths [86] or suicides [87]. Validating the information by requesting the death certificate from the practice is challenging given that GPs may not retain paper records for patients once they have died. Whilst the validity of the cause of death information in primary care has been questioned [18, 19], the quality of the data on the status or date of death information in primary care are rarely examined.

Reviews of the CPRD GOLD database have recommended linkage to other healthcare databases to validate the records [73, 88]. When this PhD started in 2011, it was not possible to validate death information in CPRD GOLD by comparing to other data sources at the individual level, however CPRD have since developed linkage programme. For deaths, two additional sources may provide additional information to corroborate the EHR data; the Hospital Episode Statistics (HES) data include information on deaths based on discharge information and the Office for National Statistics (ONS) collate the official death certificate information into a database which is considered the gold-standard data source for mortality information. There is currently no direct link between the GP practice and the General Register Office (GRO) where deaths are recorded, therefore linkage between the CPRD GOLD database and the HES and ONS death registration data has the potential to support mortality research by providing enhanced records on deaths.

Linking UK primary care with other sources of data on death

Like the Nordic databases, unique patient identifiers in the UK offer the opportunity for linking primary care data with other sources of data on death, such as the hospital records and the death registry information. Linkage between data sources is often limited to the individual countries (England, Scotland, Wales or Northern Ireland) as few data sources cover the whole nation. Prior to 2011, linkage between data sources in Scotland and Wales enabled several studies. In Scotland, the Medicines Monitoring Unit (MEMO) and Health Informatics Centre at the University of Dundee used record-linkage technology to conduct pharmacoepidemiological studies based on the population of the Tayside region [89]. Researchers have been able to use this technology alongside the Thyroid Epidemiology, Audit, and Research Study (TEARS) to study mortality in patients treated for thyroid dysfunction [90]. In Wales, the SAIL (Secure Anonymised Information Linkage) databank was established in 2006 by the Health Information Research Unit (HIRU) at the College of Medicine at Swansea University. Data from fourteen sources are currently linked, including births and deaths, demographic information, outpatient, emergency, critical care and hospitalisation data, alongside primary care data [91]. This data resource provides routinely linked disparate datasets for a range of research purposes, avoiding linkages having to be undertaken on a study specific basis. Researchers have used this linkage to study mortality following upper gastrointestinal bleeding [92].

In 2012, CPRD developed a linkage programme, which is currently limited to practices based in England, where anonymised primary care patient data from consenting GP practices have been linked individually to both the HES and ONS death registration data, alongside data on cancer, mental health and sociodemographic factors [93-96]. The linkage process uses a common methodology which enables linkage to several disparate data sets. A trusted third party, NHS Digital, holds information on unique identifiers where after linkage, all the patient identifiers are removed so that researchers only have access to anonymised information. The methodology involves an eight-step process where records from the different data sources are deterministically linked on the basis of the patient's unique NHS number, date of birth, sex and postal code of residence (postcode). As of May 2019, this included 411 practices contributing to CPRD GOLD, representing over 10 million patients, and 800 practices contributing to CPRD Aurum, representing over 25 million patients. Linkage between clinical primary care data, administrative hospital records and national mortality statistics for a large UK population has the potential to add considerable value to research studies.

Benefits and limitations of linkage

Linkage between primary care data and death registration data offers the potential to reduce misclassification by reducing the proportion of missing data and improving the accuracy of the date of death. However, this must be balanced with the reduction in cohort size and the potential for miscalculation of the denominator and follow up time.

The CPRD linkage between data sources is not always comprehensive; individuals may not consent, the general practice may decide not to participate or be based outside of England. High quality linkage is only possible for individuals with complete data on their NHS number, date of birth, sex and postcode, however, crucial identifiers such as the NHS number may not be complete [97] and therefore such patients would not be considered eligible for linkage. The proportion of patients eligible for CPRD linkage is generally around 40%, presenting a major reduction in cohort size. Researchers need to balance the impact on the study population with the increase in data available through linkage.

Each of the linked data sources will have collected data for specific periods of time (coverage period) which may or may not overlap with the time the individual was registered at the GP practice. Researchers need to consider this in the definition of their study population and follow-up time. The definition of the denominator for a study is important and there is a potential for error regarding both the individuals included and the time studied. The overlapping coverage period is the time window when data were being collected for all relevant linked data sources in the study. Potential errors arise when:

1. All practices are included regardless of eligibility for linkage
2. All patients from within an eligible practice are included regardless of individual eligibility for linkage
3. The study period is not limited to the overlapping coverage period

In scenarios 1 and 2, linked data would be missing for some individuals; this could introduce misclassification of either the exposure or the outcome, depending on the data source used to define these. Scenario 3 could lead to major misclassification of both the exposure and the outcome. In all cases, the denominator would be increased, leading to lower outcome rates. The impact would be greatest for exposures or outcomes that are defined based on the linked data, rather than the primary care record.

Some of the early studies using the linked data available through CPRD may have made these potential errors by initiating the follow-up time before the start of data collection for HES or following patients after the end of coverage for the ONS death

registration data. Two studies of multiple sclerosis analysed all potential follow-up time, including periods when one or more data source could not contribute information [98, 99]. In these examples, the studies were affected by left truncation, where it could not be accurately determined whether events had occurred before the start of the study [100]. For pharmacoepidemiology studies, it is important to be able to determine the first exposure to a drug, particularly if adverse events more commonly occur the first time an individual is exposed. Left truncation could lead to an underestimate of the event rate where subjects who tolerate the drug are overrepresented. This effect has also been coined as 'depletion of susceptibles'.

Problems with right-censoring can also occur. If the end of follow-up includes time after the end of data collection, individuals will not be at risk of the outcome during this period. Additionally, for some outcomes there is likely to be a delay between the occurrence and the recording of the event, such as the recording of a death at the GP practice, where it is possible that some deaths may be missing at the point of analysis. This has been described where censoring at death in cohort studies has an impact on disease outcomes where retrospective recording occurs [101]. It is therefore critically important to define the start and end of follow-up time, exposure and outcome that reflect the real-world events, limiting misclassification.

Thesis objectives and outline

The conclusions from an observational study of mortality can be influenced by the generalisability of the study population, the likely extend of missing information, the rarity of the outcome in the given population, and the setting of the study. Electronic healthcare records such as CPRD GOLD offer large populations in a universal healthcare setting however, there may be missing information on vital status and the date of death and delays in the recording of deaths. Linking primary care records to official death records may fill some of the gaps, however linked data may not be available for all subjects. Where only primary care information is available, researchers need to understand the limitations of using the death records since the methodology for the CPRD algorithm for estimating the date of death has not been published and the dates have not been validated. Choosing to use linked data may have an impact on the sample size for a study and adds to the complexity of defining the denominator and allocating follow-up time; researchers additionally need to consider how to deal with the potential for inconsistent data between sources. It is important to understand the strengths and limitations of the death information in primary care electronic health records alongside the opportunities and challenges of using linked data when analysing death as an outcome.

This PhD aimed to evaluate the utility of the CPRD GOLD database for studies with death as an outcome. The hypothesis being that there is sufficient recording of death dates at the GP practice to adequately censor a study population and to enable research on the fact of death with this data source.

The overall thesis objective was to evaluate the use of CPRD data for analysing death information in pharmacoepidemiological research. More specifically, the aims were:

1. To review the CPRD as a data source for pharmacoepidemiological studies of mortality
2. To conduct case studies demonstrating the use of primary care data alone for studies with a mortality outcome and when using linked data
3. To assess the impact on the study population size and follow-up time when using linked data for observational research studies
4. To evaluate the agreement between English primary care data and national death registration data in capturing vital status and the date of death

Chapter 2 provides a detailed overview on the CPRD GOLD primary care database and makes a comparison with the total Dutch population.

Chapter 3 presents examples of studies using data from CPRD to assess a mortality outcome.

Chapter 4 looks at the potential pitfalls when preparing the study population and calculating the denominator for studies using linked CPRD data.

Chapter 5 assesses how accurately vital status and the date of death are recorded at the GP practice when compared with national death registration data.

References

1. Jepsen P, Johnsen SP, Gillman MW, Sørensen HT. Interpretation of observational studies. Heart. 2004;90:956-60.
2. de Lusignan S, vanWeel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Family Practice. 2005;23:253–263.
3. Murray MD. Use of Data from Electronic Health Records for Pharmacoepidemiology. Curr Epidemiol Rep 2014;1:186-93.
4. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005;58:323-37.
5. Tierney WM, McDonald CJ. Practice databases and their uses in clinical research. Stat Med 1991;10:541–57.
6. Alonso A, Logroscino G, Jick SS, Hernán MA. Association of smoking with amyotrophic lateral sclerosis risk and survival in men and women: a prospective study. BMC Neurol. 2010;10:6.
7. Azoulay L, Schneider-Lindner V, Dell'aniello S, Schiffrin A, Suissa S. Combination therapy with sulfonylureas and metformin and the prevention of death in type 2 diabetes: a nested case-control study. Pharmacoepidemiol Drug Saf. 2010;19:335-42.
8. de Vries F, Setakis E, van Staa TP. Concomitant use of ibuprofen and paracetamol and the risk of major clinical safety outcomes. Br J Clin Pharmacol. 2010;70:429-38.
9. de Vries F, Setakis E, Zhang B, van Staa TP. Long-acting {beta}2-agonists in adult asthma and the pattern of risk of death and severe asthma outcomes: a study using the GPRD. Eur Respir J. 2010;36:494-502.
10. García Rodríguez LA, Wallander MA, Martín-Merino E, Johansson S. Heart failure, myocardial infarction, lung cancer and death in COPD patients: a UK primary care study. Respir Med. 2010;104:1691-9.
11. Khan NF, Perera R, Harper S, Rose PW. Adaptation and validation of the Charlson Index for Read/OXMIS coded databases. BMC Fam Pract. 2010;11:1.
12. McDowell SE, Coleman JJ, Evans SJ, Gill PS, Ferner RE. Laboratory monitoring and adverse patient outcomes with antihypertensive therapy in primary care. Pharmacoepidemiol Drug Saf. 2010;19:482-9.
13. Taylor CJ, Hodgkinson J, Hobbs FD. Rhythm control agents and adverse events in patients with atrial fibrillation. Int J Clin Pract. 2010;64:1069-75.
14. BMA. Confirmation and certification of death. Available at: https://www.bma.org.uk/advice/employment/gp-practices/service-provision/confirmation-and-certification-of-death. [Accessed 20 February 2018]
15. Office for National Statistics. Rolling annual death registrations by place of occurrence, England, period ending Quarter 3 (Oct to Dec) of financial year 2018 to 2019. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/009673rollingannualdeathregistrationsbyplaceofoccurrenceenglandperiodendingquarter3octtodecoffinancialyear2018to2019 [Accessed 11 March 2019]
16. Public Health England. Recent trends in place of death in England. Available at: http://www.endoflifecare-intelligence.org.uk/view?rid=893 [Accessed 7 March 2017]
17. Beaumont B, Hurwitz B. Is it possible and worth keeping track of deaths within general practice? Results of a 15 year observational study. Qual Saf Health Care. 2003;12:337-42.
18. Shah AD, Martinez C, Hemingway H. The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records. BMC Med Inform Decis Mak. 2012;12:88.

19. Hall GC. Validation of death and suicide recording on the THIN UK primary care database. Pharmacoepidemiol Drug Saf. 2009;18:120-31.
20. Crane J, Pearce N, Burgess C, *et al*. Markers of risk of asthma death or readmission in the 12 months following a hospital admission for asthma. Int J Epidemiol. 1992;21:737-44.
21. Park Y, Bateman BT, Kim DH, *et al*. Use of haloperidol versus atypical antipsychotics and risk of in-hospital death in patients with acute myocardial infarction: cohort study. BMJ. 2018;360:k1218.
22. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med. 2010;170:1989-95.
23. van der Lei J, Duisterhout JS, Westerhof HP, *et al*. The introduction of computer-based patient records in The Netherlands. Ann Intern Med. 1993;119(10):1036-41.
24. Vlug AE, van der Lei J, Mosseveld BM, *et al*. Postmarketing surveillance based on electronic patient records: the IPCI project. Methods Inf Med. 1999;38:339-44.
25. Mulder M, Baan E, Verbon A, Stricker B, Verhamme K. Trends of prescribing antimicrobial drugs for urinary tract infections in primary care in the Netherlands: a population-based cohort study. BMJ Open. 2019;9(5):e027221.
26. Engelkes M, Janssens HM, de Jongste JC, *et al*. Prescription patterns, adherence and characteristics of non-adherence in children with asthma in primary care. Pediatr Allergy Immunol. 2016;27(2):201-8.
27. Eland IA, van der Lei J, Stricker BH, Sturkenboom MJ. Incidence of priapism in the general population. Urology. 2001;57(5):970-2.
28. Stirbu-Wagner I, Dorsman SA, Visscher S, *et al*. The Netherlands Information Network of General Practice. Facts and numbers in primary care [in Dutch]. Utrecht/Nijmegen: NIVEL/IQ; 2010. [http://www.linh.nl]
29. Nielen MM, Ursum J, Schellevis FG, Korevaar JC. The validity of the diagnosis of inflammatory arthritis in a large population-based primary care database. BMC Fam Pract. 2013;14:79.
30. de Jong HJ, Vandebriel RJ, Saldi SR, *et al*. Angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers and the risk of developing rheumatoid arthritis in antihypertensive drug users. Pharmacoepidemiol Drug Saf. 2012;21(8):835-43.
31. Visser ST, Schuiling-Veninga CC, Bos JH, *et al*. The population-based prescription database IADB.nl: its development, usefulness in outcomes research and challenges. Expert Rev Pharmacoecon Outcomes Res. 2013;13(3):285-92.
32. Vroom F, van Roon EN, van den Berg PB, *et al*. Prescribing of sulfasalazine, azathioprine and methotrexate round pregnancy—a descriptive study. Pharmacoepidemiol Drug Saf. 2008;17(1):52-61.
33. de Jong J, van den Berg PB, de Vries TW, de Jong-van den Berg LT. Antibiotic drug use of children in the Netherlands from 1999 till 2005. Eur J Clin Pharmacol. 2008 Sep;64(9):913-9.
34. de Jong J, Bos JH, de Vries TW, de Jong-van den Berg LT. Use of antibiotics in rural and urban regions in The Netherlands: an observational drug utilization study. BMC Public Health. 2014 Jul 3;14:677.
35. Herings R, Pedersen L. Pharmacy-based Medical Record Linkage Systems. In: Strom BL, Kimmel SE, Hennessy S, editors. Pharmacoepidemiology. 5 ed: John Wiley & Sons, Ltd.; 2012. p. 270-86.
36. Yasmina A, de Boer A, Deneer VH, *et al*. Patterns of antiplatelet drug use after a first myocardial infarction during a 10-year period. Br J Clin Pharmacol. 2017;83(3):632-641.
37. Herk-Sukel MP, Lemmens VE, Poll-Franse LV, Herings RM, Coebergh JW. Record linkage for pharmacoepidemiological studies in cancer patients. Pharmacoepidemiol Drug Saf. 2012;21:94-103.

38. Pottegård A, Schmidt SAJ, Wallach-Kildemoes H, *et al*. Data Resource Profile: The Danish National Prescription Registry. Int J Epidemiol. 2017;46(3):798-798f.
39. Wettermark B, Zoëga H, Furu K, *et al*. The Nordic prescription databases as a resource for pharmacoepidemiological research--a literature review. Pharmacoepidemiol Drug Saf. 2013;22(7):691-9.
40. Wettermark B, Hammar N, Fored CM, *et al*. The new Swedish Prescribed Drug Register--opportunities for pharmacoepidemiological research and experience from the first six months. Pharmacoepidemiol Drug Saf. 2007;16:726-35.
41. Furu K, Wettermark B, Andersen M, *et al*. The Nordic countries as a cohort for pharmacoepidemiological research. Basic Clin Pharmacol Toxicol. 2010;106:86-94.
42. 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 Clin Pharmacol Toxicol. 2016;119(5):464-469.
43. BMA. Patient registration for GP practices. Available at: https://www.bma.org.uk/advice/employment/gp-practices/service-provision/patient-registration-for-gp-practices [Accessed 9 March 2019]
44. Kontopantelis E, Stevens RJ, Helms PJ, *et al*. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. BMJ Open. 2018;8:e020738.
45. Becher H, Kostev K, Schröder-Bernhardi D. Validity and representativeness of the "Disease Analyzer" patient database for use in pharmacoepidemiological and pharmacoeconomic studies. Int J Clin Pharmacol Ther. 2009;47:617-26.
46. Hippisley-Cox J, Stables D, Pringle M. QResearch: A new general practice database for research. Inform Prim Care 2004;12:49-50.
47. CPRD. CPRD Aurum. Available at: https://www.cprd.com/sites/default/files/CPRD%20Aurum%20FAQs_1.pdf [Accessed 2 March 2019]
48. Wolf A, Dedman D, Campbell J, *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol. 2019 Mar 11. pii: dyz034.
49. Research One. Researchers. Available at: http://www.researchone.org/researchers/ [Accessed 2 March 2019]
50. TPP. TPP and CPRD collaboration. Available at: https://www.tpp-uk.com/news/tpp-and-cprd-collaboration [Accessed 2 March 2019]
51. Lewis JD, Schinnar R, Bilker WB, *et al*. Validation studies of The Health Improvement Network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol Drug Saf 2007;16:393-401.
52. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Inform Prim Care. 2011;19:251-5.
53. Hall GC, Luscombe DK, Walker SR. Post marketing surveillance using a computerised general practice database. Pharmaceutical Med 1988;2:345–51.
54. Walley T, Mantgani A. The UK General Practice Research Database. Lancet. 1997;350:1097-9.
55. Lawson DH, Sherman V, Hollowell J. The General Practice Research Database. Scientific and Ethical Advisory Group. QJM. 1998;91:445-52.
56. The General Practice Research Database. GPRD recording guidelines for Vision users. London: Crown Publishing, 2004.
57. Benson T. The history of the Read Codes: the inaugural James Read Memorial Lecture 2011. Inform Prim Care. 2011;19:173-82.
58. O'Neil M, Payne C, Read J. Read Codes Version 3: a user led terminology. Methods Inf Med. 1995;34:187-92.

59. Wood L, Martinez C. The general practice research database: role in pharmacovigilance. Drug Saf. 2004;27:871-81. Review.
60. Wood L, Coulson R. Revitalizing the General Practice Research Database: plans, challenges, and opportunities. Pharmacoepidemiol Drug Saf. 2001;10:379-83.
61. QOF. Available at: https://qof.digital.nhs.uk/ [Accessed 18 March 2019]
62. Smeeth L, Cook C, Fombonne E, *et al*. MMR vaccination and pervasive developmental disorders: a case–control study. Lancet 2004;364: 963–9.
63. Shephard EA, Stapley S, Neal RD, *et al*. Clinical features of bladder cancer in primary care. Br J Gen Pract. 2012 Sep;62(602):e598-604.
64. Bhaskaran K, Douglas I, Forbes H, *et al*. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. Lancet. 2014;384:755-65.
65. Lalmohamed A, MacGregor AJ, de Vries F, Leufkens HG, van Staa TP. Patterns of risk of cancer in patients with metal-on-metal hip replacements versus other bearing surface types: a record linkage study between a prospective joint registry and general practice electronic health records in England. PLoS One. 2013;8:e65891.
66. Driessen JH, van Onzenoort HA, Henry RM, *et al*. Use of dipeptidyl peptidase-4 inhibitors for type 2 diabetes mellitus and risk of fracture. Bone. 2014;68:124-30.
67. de Vries F, Souverein PC, Cooper C, Leufkens HG, van Staa TP. Use of beta-blockers and the risk of hip/femur fracture in the United Kingdom and The Netherlands. Calcif Tissue Int. 2007;80:69-75.
68. Hall GC, Sauer B, Bourke A, *et al*. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2012;21:1-10.
69. Crooks CJ, Card TR, West J. Excess long-term mortality following non-variceal upper gastrointestinal bleeding: a population-based cohort study. PLoS Med. 2013;10:e1001437.
70. Cotton SJ, Belcher J, Rose P, K Jagadeesan S, Neal RD. The risk of a subsequent cancer diagnosis after herpes zoster infection: primary care database study. Br J Cancer. 2013;108:721-6.
71. Hamilton WT, Round AP, Sharp D, Peters TJ. The quality of record keeping in primary care: a comparison of computerised, paper and hybrid systems. Br J Gen Pract. 2003;53:929-33.
72. Hoffmann F, Andersohn F, Giersiepen K, *et al*. Validation of secondary data. Strengths and limitations. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2008;51:1118-26.
73. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol 2010;69:4-14.
74. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract. 2010;60:e128-36.
75. Hennessy S, Leonard CE, Palumbo CM, *et al*. Diagnostic codes for sudden cardiac death and ventricular arrhythmia functioned poorly to identify outpatient events in EPIC's General Practice Research Database. Pharmacoepidemiol Drug Saf. 2008;17:1131-6.
76. Hersh WR, Weiner MG, Embi PJ, *et al*. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-7.
77. Wagstaff R, Berlin A, Stacy R, Spencer J, Bhopal RA. Information about patients' deaths: general practitioners' current practice and views on receiving a death register. Br J Gen Pract. 1994;44:315-6.

78. INPS Ltd. 2016. Consultation Manager - Data Entry General Points. Available at: http://www.inps4.co.uk/sites/default/files/consultation_manager_-_data_entry_general_points.pdf [Accessed 14 March 2019]
79. Azoulay L, Schneider-Lindner V, Dell'aniello S, *et al*. Thiazolidinediones and the risk of incident strokes in patients with type 2 diabetes: a nested case-control study. Pharmacoepidemiol Drug Saf. 2010;19:343-50.
80. Becker C, Brobert GP, Johansson S, Jick SS, Meier CR. Risk of incident depression in patients with Parkinson disease in the UK. Eur J Neurol. 2011;18:448-53.
81. Breart G, Cooper C, Meyer O, Speirs C, Deltour N, Reginster JY. Osteoporosis and venous thromboembolism: a retrospective cohort study in the UK General Practice Research Database. Osteoporos Int. 2010;21:1181-7.
82. Charlton RA, Weil JG, Cunnington MC, de Vries CS. Identifying major congenital malformations in the UK General Practice Research Database (GPRD): a study reporting on the sensitivity and added value of photocopied medical records and free text in the GPRD. Drug Saf. 2010;33:741-50.
83. Schneider C, Bothner U, Jick SS, Meier CR. Chronic obstructive pulmonary disease and the risk of cardiovascular diseases. Eur J Epidemiol. 2010;25:253-60.
84. Schneider C, Jick SS, Bothner U, Meier CR. Cancer risk in patients with chronic obstructive pulmonary disease. Pragmat Obs Res. 2010;1:15-23. eCollection 2010.
85. Shah A.D, Martinez C. A Comparison of the Cause of Death Recorded in the General Practice Research Database with National Mortality Statistics in England and Wales. In: Abstracts of the 20th International Conference on Pharmacoepidemiology (ICPE) & Therapeutic Risk Management. Bordeaux, France, 22 - 25 August, 2004. Pharmacoepidemiol Drug Saf. 2004 Jul;13: S1-S349.
86. Murray-Thomas T, Jones ME, Patel D, *et al*. Risk of mortality (including sudden cardiac death) and major cardiovascular events in atypical and typical antipsychotic users: a study with the general practice research database. Cardiovasc Psychiatry Neurol. 2013;2013:247486.
87. Thomas KH, Davies N, Metcalfe C, *et al*. Validation of suicide and self-harm records in the Clinical Practice Research Datalink. Br J Clin Pharmacol. 2013;76:145-57.
88. García Rodríguez LA, Pérez Gutthann S. Use of the UK General Practice Research Database for pharmacoepidemiology. Br J Clin Pharmacol. 1998;45:419-25.
89. Evans JM, MacDonald TM. Record-linkage for pharmacovigilance in Scotland. Br J Clin Pharmacol. 1999;47,105–10.
90. Flynn RW, Macdonald TM, Jung RT, *et al*. Mortality and vascular outcomes in patients treated for thyroid dysfunction. J Clin Endocrinol Metab. 2006;91:2159-64.
91. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, Brown G, Leake K. The SAIL databank: linking multiple health and social care datasets. BMC Medical Informatics and Decision Making, 2009;9:3.
92. Roberts SE, Button LA, Williams JG. Prognosis following upper gastrointestinal bleeding. PLoS One. 2012;7(12):e49507.
93. Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK primary care data resource. Ther Adv Drug Saf. 2012;3(2):89-99.
94. Gallagher AM, Puri S, van Staa TP. Linkage of the General Practice Research Database (GPRD) with Other Data Sources. Pharmacoepidemiol Drug Saf. 2011;20:S230-231.
95. Eaton S, Setakis E, Williams TW, van Staa TP. Linking Primary Care Data (UK GPRD) to Hospital Records (HES). Pharmacoepidemiol Drug Saf. 2010;19:S195.

96. Setakis E, Puri S, Gallagher A, *et al*. Linking UK Death Certificate Data to UK Primary Care Data (General Practice Research Database). Pharmacoepidemiol Drug Saf. 2010;19:S310-311.
97. Hippisley-Cox J. Validity and completeness of the NHS Number in primary and secondary care Electronic data in England 1991-2013. 2015. Available at https://www.qresearch.org/publications/downloads/ http://eprints.nottingham.ac.uk/3153/3/QResearchReport2013.pdf
98. Bazelier MT, Mueller-Schotte S, Leufkens HG, *et al*. Risk of cataract and glaucoma in patients with multiple sclerosis. Mult Scler. 2012;18:628-38.
99. Bazelier MT, van Staa TP, Uitdehaag BM, *et al*. A simple score for estimating the long-term risk of fracture in patients with multiple sclerosis. Neurology. 2012;79:922-8.
100. Hazelbag CM, Klungel OH, van Staa TP, *et al*. Left truncation results in substantial bias of the relation between time-dependent exposures and adverse events. Ann Epidemiol. 2015;25:590-6.
101. Binder N, Blümle A, Balmford J, *et al*. Cohort studies were found to be frequently biased by missing disease information due to death. J Clin Epidemiol. 2019;105:68-79.

Chapter 2:

CPRD as a data source

Chapter 2.1

Data Resource Profile: Clinical Practice Research Datalink (CPRD)

Emily Herrett, Arlene M Gallagher, Krishnan Bhaskaran,
Harriet Forbes, Rohini Mathur, Tjeerd van Staa, Liam Smeeth

Int J Epidemiol. 2015;44(3):827-36

Summary

The Clinical Practice Research Datalink (CPRD) is an ongoing primary care database of anonymised medical records from general practitioners, with coverage of over 11.3 million patients from 674 practices in the UK. With 4.4 million active (alive, currently registered) patients meeting quality criteria, approximately 6.9% of the UK population are included and patients are broadly representative of the UK general population in terms of age, sex and ethnicity. General Practitioners are the gatekeepers of primary care and specialist referrals in the UK. The CPRD primary care database is therefore a rich source of health data for research, including data on demographics, symptoms, tests, diagnoses, therapies, health-related behaviours and referrals to secondary care. For over half of patients, linkage with datasets from secondary care, disease-specific cohorts and mortality records enhance the range of data available for research. The CPRD is very widely used internationally for epidemiological research and has been used to produce over 1000 research studies, published in peer reviewed journals across a broad range of health outcomes. However, researchers must be aware of the complexity of routinely collected electronic health records, including ways to manage variable completeness, misclassification and development of disease definitions for research.

Data resource basics

UK primary care data for research

Over 98% of the UK population are registered with a primary care General Practitioner (GP) [1] and under the National Health Service (NHS), visits to the GP are free of charge. The GP is the gatekeeper of care in the UK National Health Service. GPs act as the first point of contact for any non-emergency health-related issues, which may then be managed within primary care, and/or referred to secondary care as necessary. Secondary care teams also feed back information to GPs about their patients, including key diagnoses. Patient data are routinely recorded onto computers by practice staff, against a unique patient NHS number. These facets of UK primary care provide good capture of health information in a longitudinal electronic health record.

The Clinical Practice Research Datalink (CPRD)

The CPRD harnesses general practice data and produces a primary care dataset, which is one of the largest databases of longitudinal medical records from primary care in the world (Table 1). Established in London in 1987, the small Value Added Medical Products (VAMP) dataset grew to become the General Practice Research Database (GPRD) in 1993, [2, 3] before expanding to become CPRD in 2012. The CPRD collates routinely collected anonymised electronic health record data from general practices who have agreed at a practice level to provide data on a monthly basis. All patients registered with the participating practices are included in the dataset, unless they have individually requested to opt out of data sharing, by asking their GP to amend their registration details on the system to disable the extraction of their data.

Data linkage

A subset of English practices (currently 75%, representing 58% of all UK CPRD practices) have consented to participate in the CPRD linkage programme and have provided patient level information. Patient level data from consenting practices are linked via a trusted third party (the Health and Social Care Information Centre [4]) to other existing data sources. Established linkages include Hospital Episode Statistics [5] (hospitalisation data), Office for National Statistics [6] (mortality data including causes of death), Index of Multiple Deprivation and Townsend scores (deprivation data) and disease registries including the National Cancer Intelligence Network [7], and the Myocardial Ischaemia National Audit Project [8] (Table S1, supporting information). Other linkages are planned (see CPRD website [9]) and researchers can make requests for bespoke linkage for individual studies.

Uses for observational research and interventional research

Subject to the appropriate data governance and approvals, the CPRD can supply primary care and linked patient data to researchers in the UK and internationally. Through the CPRD, researchers can approach practices and patients to take part in biosample collection studies or trials. The feasibility of this work has been tested: patients from CPRD have been recruited to a pharmacogenetic study of statin-induced myopathy [10, 11], practices have been recruited to cluster randomised trials [12, 13] and patients have been recruited to pragmatic point-of-care randomised trials [14]. The electronic health record data can be used alongside the study data to provide a full clinical picture for the recruited patients.

Ethics

The CPRD has broad National Research Ethics Service Committee (NRES) ethics approval for purely observational research using the primary care data and established data linkages. Other uses of CPRD data may require separate ethical approval. This is likely if there is any specific patient involvement in the study; for example, if the researcher wishes to ask patients to complete a questionnaire for Patient Reported Outcomes, or to conduct an interventional trial among CPRD patients.

Data governance, practice and patient confidentiality

The CPRD strives to operate within UK and European Laws to protect confidentiality. Governance requirements to protect patient confidentiality where patient consent has not been obtained are respected by ensuring patient identifiers are held separately from the clinical data and that there is separation between researchers with access to identifiable information from the primary study and those using CPRD data.

Funding sources

The CPRD is a joint venture from the Medicines and Healthcare Regulatory Agency (MHRA) and the National Institute for Health Research (NIHR). The CPRD is owned by the UK Department of Health and operates within the MHRA. The CPRD has received funding for studies from the MHRA, Wellcome Trust, Medical Research Council, NIHR Health Technology Assessment programme, Innovative Medicines Initiative, UK Department of Health, Technology Strategy Board, Seventh Framework Programme EU, various universities, contract research organizations and pharmaceutical companies.

Data resource area and population coverage

Figure 1 describes the population coverage of CPRD primary care data across England, Wales, Scotland and Northern Ireland. At the midyear date of 2nd July 2013,

the dataset held information on 11.3 million patients, who were deemed acceptable for research based on data quality checks (Appendix S1, supporting information, and described below). The population of active patients (alive and currently registered) on 2nd July 2013 was 4.4 million, representing 6.9% of the UK population (based on the UK 2013 midyear population of 64.1 million). The remaining 6.9 million records represent inactive patients, who have died or are no longer registered with a participating practice. Patient numbers by age, sex, deprivation, ethnicity and region are described in Table 2.

Frequency of data collection

Data collection happens as part of normal clinical care of patients in participating practices on a daily basis. The frequency of data recording is determined by patient need and varies by age, sex and underlying morbidity. Patients are included in the primary care dataset from their first until their last contact with the participating practice. Data are collected by practices and usually uploaded to the CPRD secure servers on a monthly basis. The date of last data collection corresponds to the date of the last data upload from each practice. Monthly builds of the primary care dataset are generated and made available for researchers to use.

Measures

Practice and patient data

The database structure broadly separates information into clinical, referral, immunisation, test and therapy data (Table S1, supporting information). Data are recorded against practice and patient pseudo-identifiers. At the practice level, geographic region is recorded by the CPRD as one of 10 regions in England, with Wales, Scotland and Northern Ireland as separate regions (Figure 1); a practice-level deprivation score is also calculated based on practice lower super output area.

All general practice encounters are recorded electronically and practitioners are encouraged to make these records available for research. Data are collected on demographic information, prescription details, clinical events (symptoms, diagnoses), preventative care provided, tests, immunisations, specialist referrals, hospital admissions and their major outcomes, and detail relating to death (Table S2, supporting information).

All entries to a patient record are considered as ‘consultations’, not all of which will involve a face to face encounter. Within a consultation multiple ‘events’ may be recorded, each with an associated date (Figure 2).

Data are largely recorded by general practice staff using version 2 Read codes, a hierarchical clinical classification system containing over 96,000 codes [15]. For example, during a consultation, a GP, nurse, other healthcare professional, practice manager or administrator may enter a number of Read codes to describe a patient's condition (e.g. lifestyle measures such as smoking status, symptoms, past medical history, diagnoses, tests performed such as blood pressure measurement, and therapies offered). Numerical data on additional clinical measures (e.g. height, weight, blood pressure, alcohol intake) can also be recorded during consultations. Prescriptions issued by the GP are automatically recorded with a product name and British National Formulary code, alongside the dosage instructions and quantity. Results of laboratory tests ordered by the GP are commonly added to the patient record via electronic links to laboratories. Data fed back to the GP from other sources may also be entered into the patient record by practice staff; this might include information from secondary care such as key diagnoses, discharge data from hospitals, or follow-up information from specialist clinics.

The GP is also able to make additional uncoded notes and observations about patients as free text. This often contains identifiable information and is not part of the standard database available to researchers.

Data resource use

Data from the CPRD (or formerly the GPRD or VAMP) have been used in the UK and internationally [16] to produce close to 2000 research reports, with over 1000 published in peer reviewed journals, across all major therapeutic areas. A bibliography is maintained by the CPRD and is available online [17]. These publications cover a range of health-related research topics including pharmacoepidemiology, comparative effectiveness research, health services research, assessments of temporal trends in disease incidence, health economics, prognosis research, classical risk factor epidemiology, and more recently randomised controlled trials [12, 18]. Publications to date include studies showing the absence of an association between measles, mumps and rubella (MMR) vaccine and autism [19], cardiovascular risk after acute infection [20], the lower risk of dementia associated with statin use [21], the risk of myocardial infarction in patients with psoriasis [22], the use of oral corticosteroids and fracture risk [23], and the association between body mass index and cancer [24].

Strengths

The strengths of the CPRD data as a research resource lie in the breadth of coverage, size, long term follow-up, representativeness, and data quality.

Breadth of data

The CPRD primary care dataset is one of few large, ongoing databases that includes data on morbidity and lifestyle variables and with a linkage to secondary care and mortality data.

Size and long-term follow-up

A key strength of this database is its size; the CPRD hold data from 674 practices and includes over 79 million person-years of follow-up (on 2nd July 2013, January 2014 dataset). This allows epidemiological associations to be investigated in more detail and estimated with a higher level of statistical precision than is possible with smaller data sources, which is of particular importance for the study of rare exposures and diseases [25, 26]. For individual patients, there is a median prospective follow-up of 9.4 years for active patients (interquartile range (IQR) 3.4-13.9) and 5.1 years (IQR 1.8-11.1 years) (Table 2) overall, enabling research into diseases with long latency and the study of long-term outcomes [27-29].

Representativeness

When compared to the UK census in 2011 [30], CPRD patients are broadly representative of the UK population in terms of age and sex (Figure 3). Patients are also comparable to the UK census in terms of ethnicity [31], and comparable to the Health Survey for England for body mass index distribution in most patient subgroups [32]. However, the CPRD may not be representative of all practices in the UK based on geography and size [33].

Data quality

Aspects of data quality in English general practice are enhanced by the Quality and Outcomes Framework [34], an incentive payment programme for GPs, which encourages recording of key data items (for example smoking status and the delivery of services to key patient groups). The Quality and Outcomes Framework was introduced in 2004 and completeness of many variables showed subsequent improvement in recording (Figure 4 and Figures S1-S2, supporting information).
Validation of the CPRD has shown high positive predictive value of some diagnoses and, where evaluated, comparisons of incidence to other UK data sources are also broadly similar [35-38]. However, reporting of validation studies was often too poor to

permit a clear interpretation, and the majority of studies focussed on positive predictive value rather than sensitivity or specificity [39].

The quality of primary care data is variable because data are entered by GPs during routine consultations, not for the purpose of research. Researchers must therefore undertake comprehensive data quality checks before undertaking a study. The CPRD provides two sets of data quality criteria: acceptability for patients and up to standard (UTS) time for practices. These criteria do not ensure data quality, but the CPRD recommend that these measures are used as a first step to selecting research quality patients and periods of quality data recording. The acceptable patient metric is based on registration status, recording of events in the patient record, and valid age and gender. The UTS date is a practice-based quality metric based on the continuity of recording and the number of recorded deaths. The UTS date is calculated for each participating practice, corresponding to the latest date at which practices meet these minimum quality criteria (Appendix S1, supporting information). The figures given in this paper reflect data for patients labelled as acceptable and who have at least one day of follow-up that is 'up to standard'. Research into data quality has shown that, despite these criteria, there were large variations in inter-practice recording of data [40].

Weaknesses

Missing data

The variability in completeness of data across patients and across time requires careful consideration; restricting to those with complete data may result in biased analyses and imputation may not be a straightforward approach because the patterns of missingness are complex. For example, body mass index may be recorded more frequently in patients with a health issue, and blood pressure more frequently in women of reproductive age and those with existing cardiovascular disease. Complex algorithms are often required to deal with missingness, to resolve discrepancies in measures between consultations, and to decide whether historical measurements, for example of body mass index, blood pressure or smoking status, are still appropriate to a patient's disease risk much later in follow-up [32].
An additional complexity of primary care data is that the absence of a Read code for disease must be interpreted as an absence of the disease itself, so while positive predictive value tends to be high [39], sensitivity may be lower. This potential misclassification arises partly due to patients failing to present to the GP with disease, and also from variations between GPs in coding diagnoses in the patient electronic record; if GPs enter information as free text, researchers will miss valuable information. The extent of misclassification may vary between diseases [39].

Definitions

There are not generally standardised definitions for diagnoses and other details, so Read code lists and algorithms need to be developed for each study to identify exposures and outcomes of interest. This may lead to inconsistent definitions (and therefore results) between studies using the same data.

Information from secondary care

General practices receive information about patient contacts with secondary care but this information must be manually entered into the patient record. Therefore, details about hospital admissions (dates, diagnoses, tests performed, length of stay) may be incomplete.

Data not captured

Some aspects of health may be recorded very infrequently or not at all, for example level of social support, number of people in a household, over the counter medication use, prescriptions in secondary care, prescriptions filled, and adherence to treatments. There are also certain patient groups that are missing from primary care records, such as prisoners, private patients, some residential homes and the homeless.

Data Resource access

Access to patient level data is provided by the CPRD for health research purposes and is dependent on approval of a study protocol by the MHRA Independent Scientific Advisory Committee (ISAC).

Researchers intending to use the data should be aware that the CPRD data files contain millions of rows of data, requiring extensive data management and an in-depth understanding of the way the data are input and stored.

The CPRD provide data dictionaries and coding dictionaries to researchers, and guidance on creating code lists is available to help identify codes of interest [41]. Read code repositories for electronic health record research are also now available [42, 43].

Details about ISAC applications and data costs are available on the CPRD website, and any other queries can be directed to the CPRD Knowledge Centre (kc@cprd.com) [9].

Key messages

- CPRD data have been extensively used for observational research. For example, the data were used to show there was no association between MMR vaccine and autism, and to show an association between oral corticosteroid use and increased risk of fractures;

- The CPRD has a large UK dataset bringing together longitudinal primary care medical records from participating practices. Over half of CPRD patients are eligible for linkage to additional datasets, including hospital data, national cancer registration data, and national mortality records;

- Quality of some data is driven by the Quality and Outcomes Framework in the UK, and data are also monitored by CPRD internal processes. Analyses described in this paper show that active (alive, currently registered) CPRD patients are representative of the UK population in terms of age, sex and ethnicity;

- CPRD data originate from routine clinical practice, and their use for epidemiological studies typically requires extensive data processing and an understanding of the way the data are originally recorded and stored.

Table 1: Key details about the Clinical Practice Research Datalink

Countries participating	UK - England, Wales, Scotland and Northern Ireland
Who is included?	Patients registered at general practices that contribute data to CPRD, who have not dissented from secondary use of GP patient identifiable data
What is recorded?	Demographics, diagnoses, symptoms, signs, prescriptions, referrals, immunisations, behavioural factors, tests
Period of data collection	1987 to present
	Average duration of follow-up 5.1 years
Funding source	CPRD has received funding from the MHRA, Wellcome Trust, Medical Research Council, NIHR Health Technology Assessment programme, Innovative Medicines Initiative, UK Department of Health, Technology Strategy Board, Seventh Framework Programme EU, various universities, contract research organizations and pharmaceutical companies

UK: United Kingdom; CPRD: Clinical Practice Research Datalink; MHRA: Medicines and Healthcare products Regulatory Agency; NIHR: National Institute for Health Research; EU: European Union.

Table 2: Demographic characteristics of acceptable CPRD patients (January 2014 dataset build), and the subset of those active on 2nd July 2013

	All patients		Active	
Number of patients	11,299,221		4,425,016	
Men, *n* (%)	5,478,715	(48.5)	2,183,161	(49.3)
Women, *n* (%)	5,820,506	(51.5)	2,241,855	(50.7)
Age in 2013, n (%) (years)				
<18	-		892,670	(20.2)
18-64	-		2,733,657	(61.8)
65+	-		798,689	(18.1)
Region, *n* (%)				
North East	184,753	(1.6)	67,639	(1.5)
North West	1,257,846	(11.1)	523,356	(11.8)
Yorkshire & The Humber	441,933	(3.9)	48,480	(1.1)
East Midlands	446,799	(4)	29,954	(0.7)
West Midlands	943,011	(8.4)	394,115	(8.9)
East of England	1,117,235	(9.9)	306,538	(6.9)
South West	943,295	(8.4)	377,821	(8.5)
South Central	1,236,351	(10.9)	544,979	(12.3)
London	1,532,066	(13.6)	600,824	(13.6)
South East Coast	1,130,468	(10)	474,593	(10.7)
Northern Ireland	275,640	(2.4)	153,576	(3.5)
Scotland	960,121	(8.5)	499,969	(11.3)
Wales	829,703	(7.3)	403,172	(9.1)
Duration of follow-up				
(median years, IQR)[a]	5.1	(1.8-11.1)	9.4	(3.4-13.9)

Active patients are alive and currently registered on 2nd July 2013.
[a]Includes only up to standard follow-up
CPRD: Clinical Practice Research Datalink; IQR: Interquartile range.

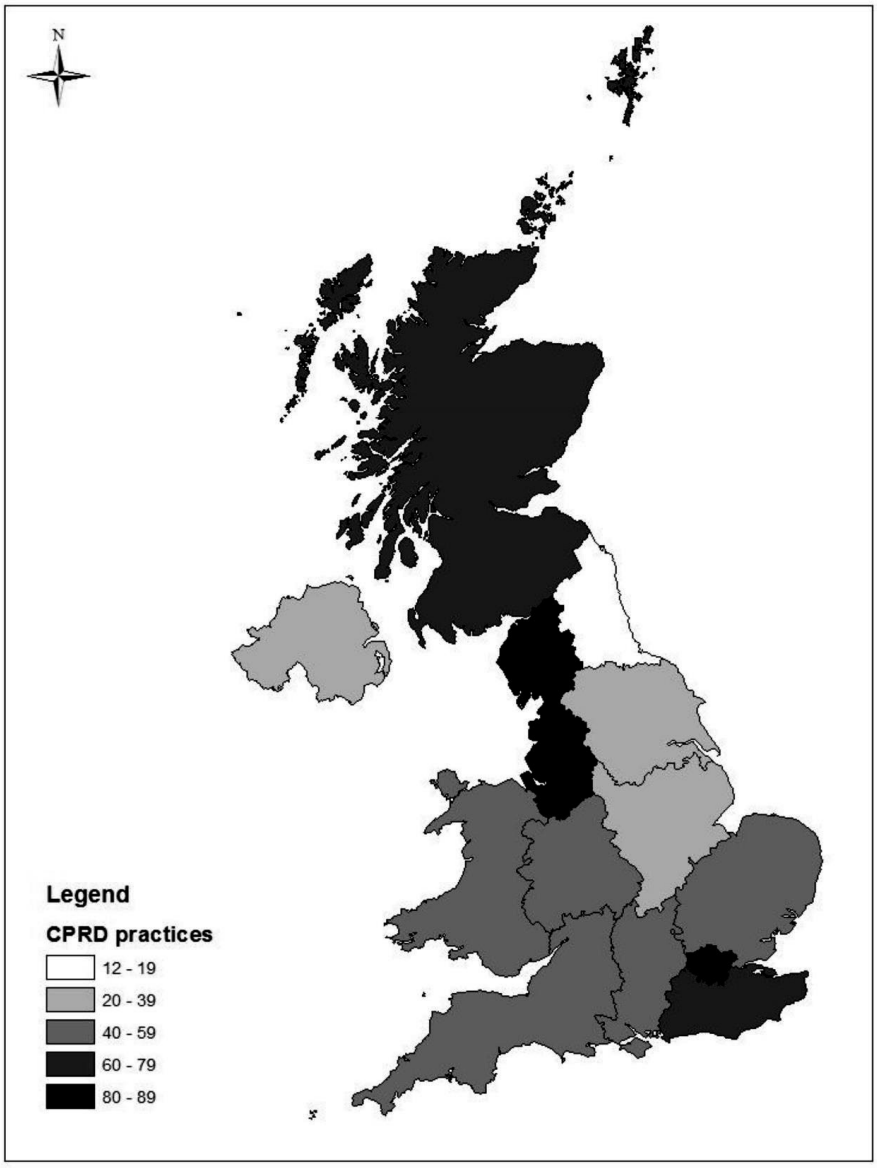


Figure 1: Distribution of 674 CPRD practices by region in England, and in Wales, Scotland and Northern Ireland.

Note: practices mapped are those contributing up to standard data to the dataset on 2nd July 2013, based on the January 2014 dataset build.

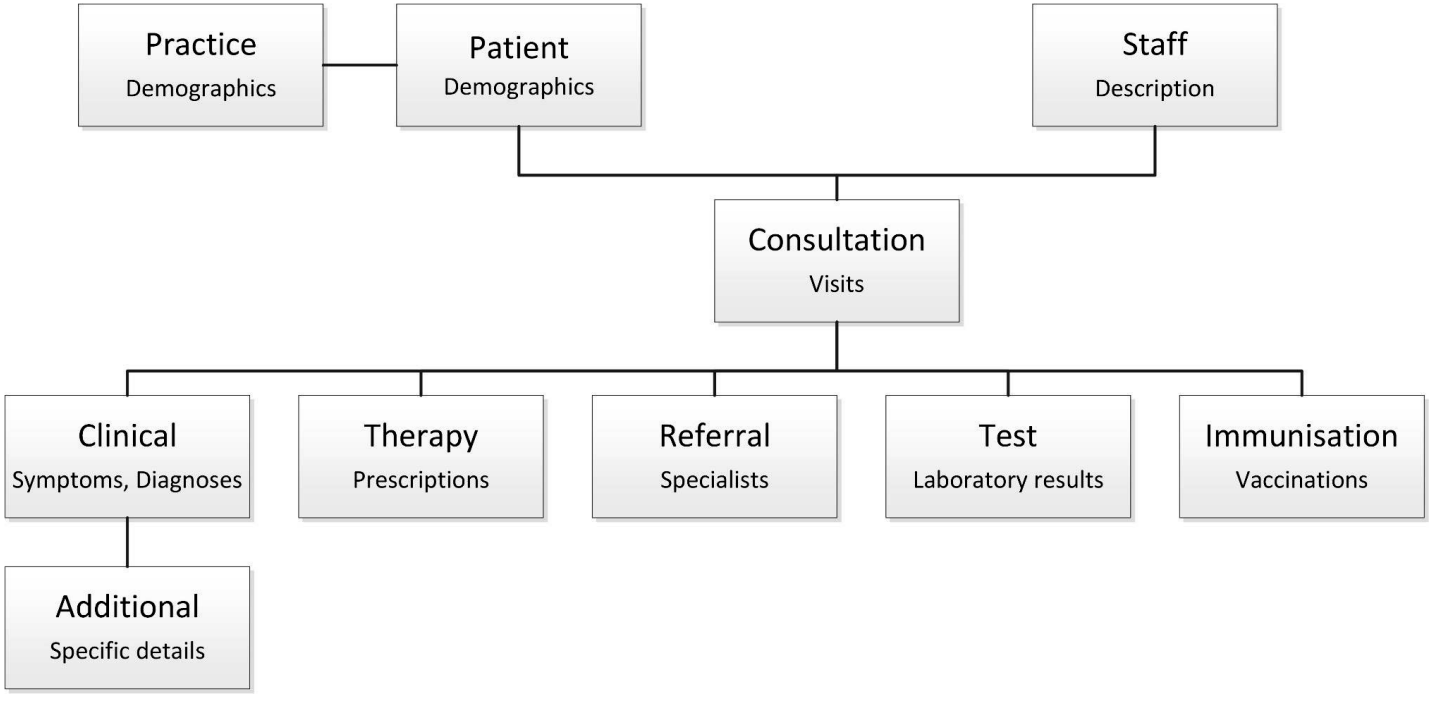


Figure 2: Example of dataset structure.

Note: patients consult with practice staff, where clinical, therapy, referral, test and immunisation information is coded in the medical record.

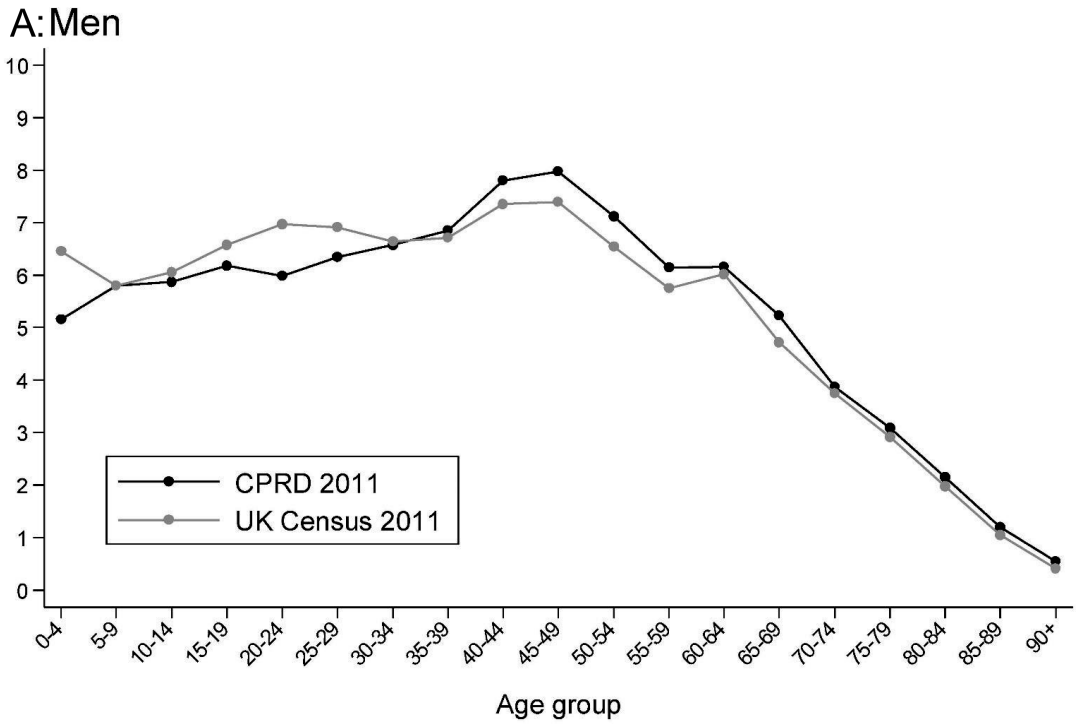


Figure 3: Age distribution of the CPRD primary care data on 27th March 2011 compared to UK Census data 2011 (A: Men; B: Women).

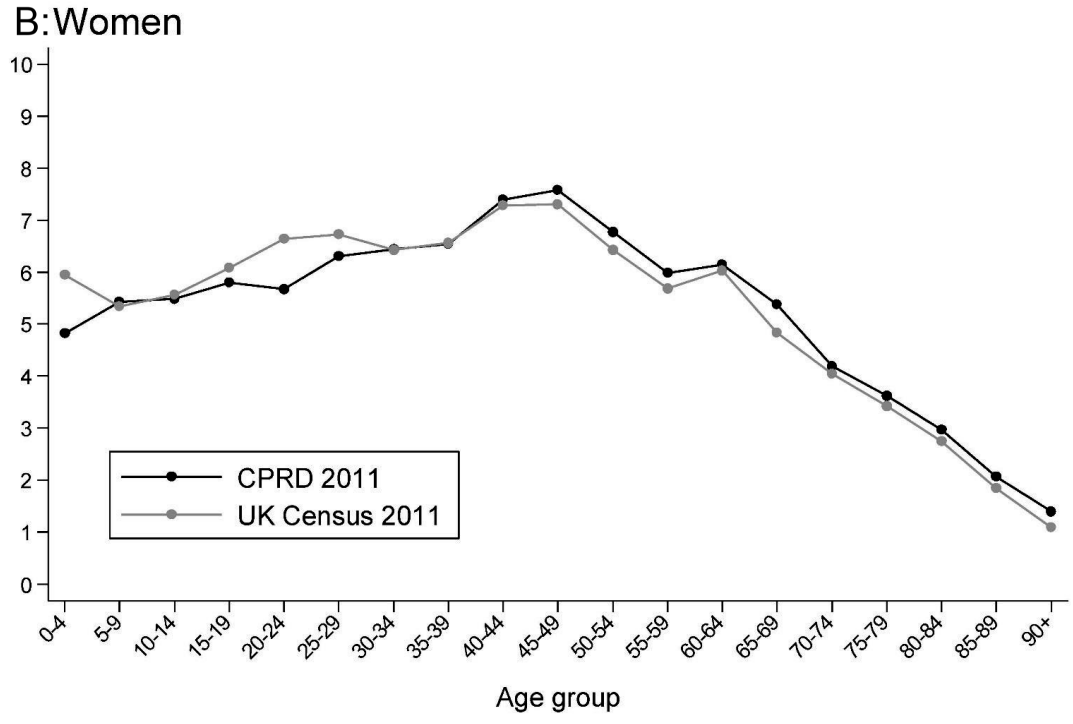


Figure 3 (cont.): Age distribution of the CPRD primary care data on 27th March 2011 compared to UK Census data 2011 (A: Men; B: Women).

Note: These data are based on a one million patient sample of primary care data from the CPRD. All patients are acceptable.
CPRD: Clinical Practice Research Datalink; UK: United Kingdom.

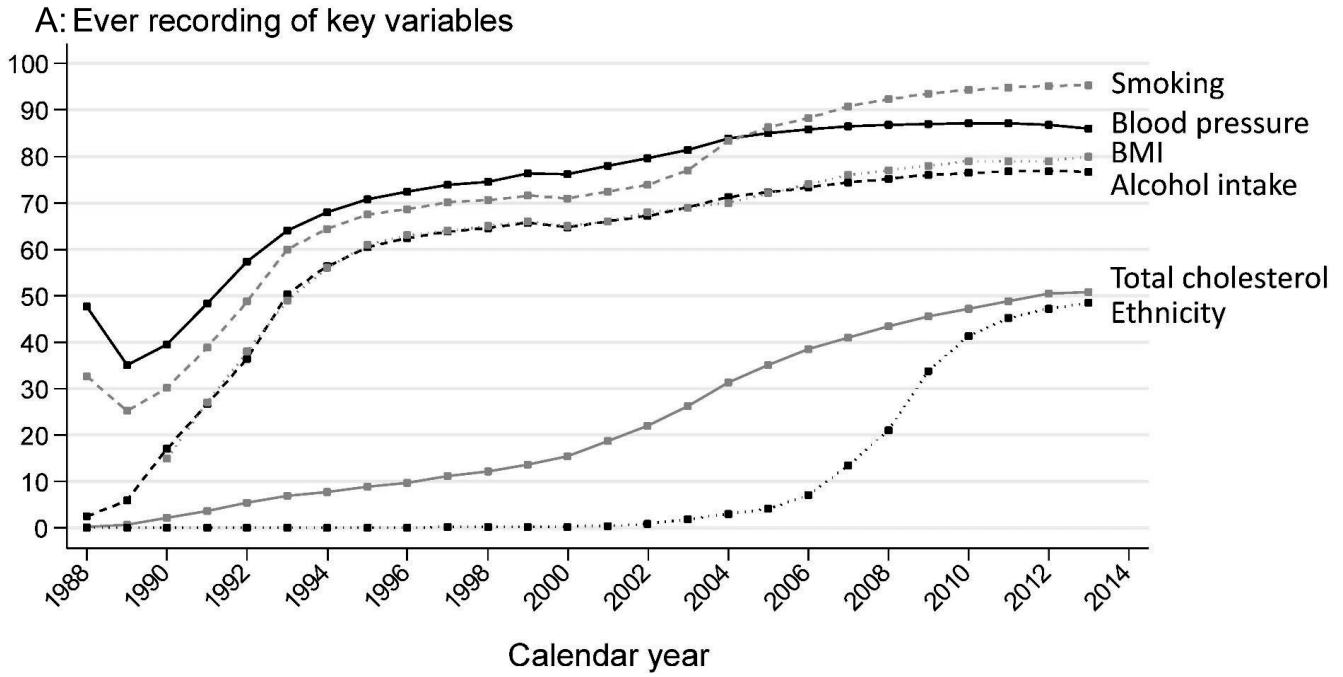


Figure 4: Recording of key lifestyle and demographic variables by calendar year
(A: ever recorded in patient follow-up; B: recorded in the past three years of patient follow-up).

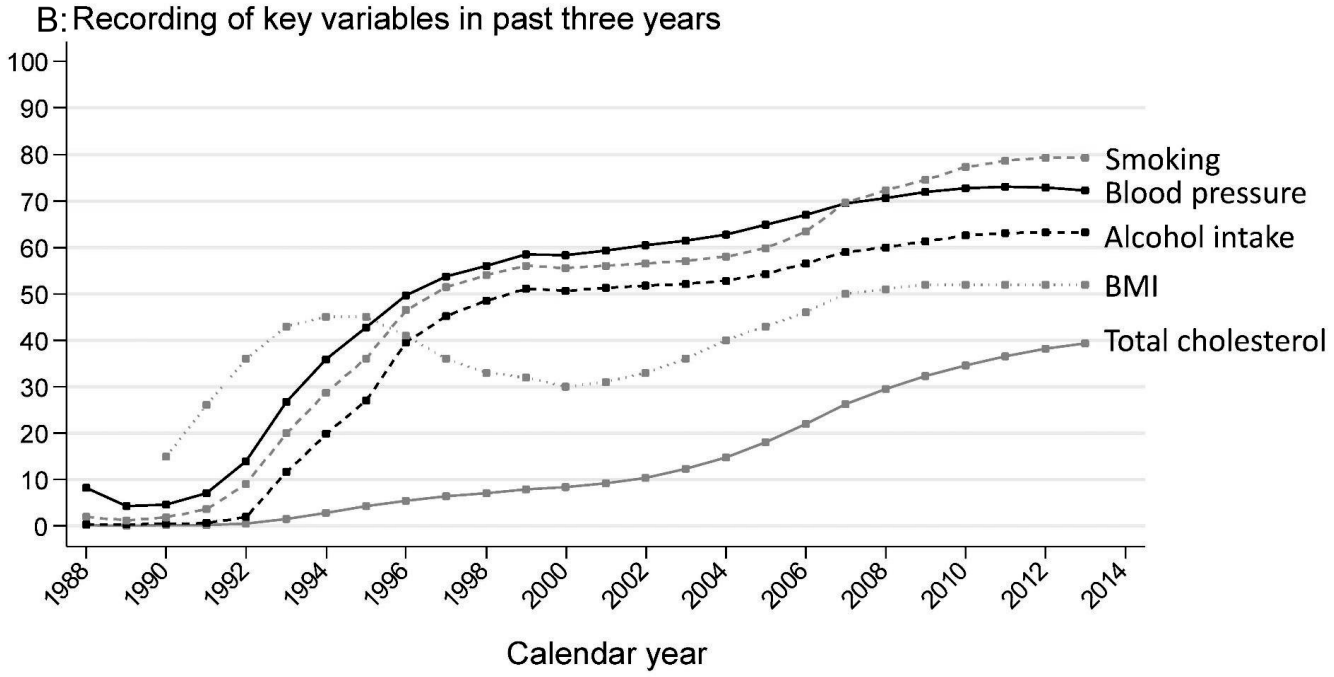


Figure 4 (cont.): Recording of key lifestyle and demographic variables by calendar year
(A: ever recorded in patient follow-up, B: recorded in the past three years of patient follow-up).

Note: These data are based on a one million patient sample of primary care data from the CPRD. All patients are acceptable. Adults aged 18+ were included. Records outside of UTS were included. Denominators are mid-year registered populations. Total cholesterol has poorer completeness as this would not be routinely recorded and would require a clinical indication. Completeness of ethnicity recording for new registrants after 2004 approached 70% in 2011.[31]
BMI: Body Mass Index; CPRD: Clinical Practice Research Datalink; UTS: Period of up-to-standard data recording.

References

1. Health and Social Care Information Centre. Attribution Data Set GP-Registered Populations Scaled to ONS Population Estimates - 2011. Available at: http://www.hscic.gov.uk/catalogue/PUB05054. [Accessed 6 August 2014]
2. Kousoulis AA, Rafi I, de Lusignan S. The CPRD and the RCGP: building on research success by enhancing benefits for patients and practices. Br J Gen Pract. 2015;65:54-5.
3. Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. Ther Adv Drug Saf. 2012;3:89-99.
4. Health and Social Care Information Centre (HSCIC). Available at: http://www.hscic.gov.uk/ [Accessed 6 August 2014]
5. Hospital Episode Statistics. Available at: http://www.hscic.gov.uk/hes [Accessed 21 January 2015]
6. Office for National Statistics, Mortality statistics: Metadata. 2014: Cardiff.
7. National Cancer Intelligence Network. Available at: http://www.ncin.org.uk/home [Accessed 23 January 2015]
8. Herrett E, Smeeth L, Walker L, Weston C. The Myocardial Ischaemia National Audit Project (MINAP). Heart. 2010;96:1264-67.
9. The Clinical Practice Research Datalink (CPRD). Available at: www.cprd.com [Accessed 6 August 2014]
10. Carr DF, O'Meara H, Jorgensen AL, Campbell J, Hobbs M, McCann G, van Staa T, Pirmohamed M. SLCO1B1 genetic variant associated with statin-induced myopathy: a proof-of-concept study using the clinical practice research datalink. Clin Pharmacol Ther. 2013;94:695-701.
11. O'Meara H, Carr DF, Evely J, Hobbs M, McCann G, van Staa T, Pirmohamed M. Electronic health records for biological sample collection: feasibility study of statin-induced myopathy using the Clinical Practice Research Datalink. Br J Clin Pharmacol. 2014;77:831-38.
12. Horspool MJ, Julious SA, Boote J, Bradburn MJ, Cooper CL, Davis S, Elphick H, Norman P, Smithson WH, van Staa T. Preventing and lessening exacerbations of asthma in school-age children associated with a new term (PLEASANT): study protocol for a cluster randomised control trial. Trials. 2013;14:297.
13. Gulliford MC, van Staa TP, McDermott L, McCann G, Charlton J, Dregan A; eCRT Research Team. Cluster randomized trials utilizing primary care electronic health records: methodological issues in design, conduct, and analysis (eCRT Study). Trials. 2014;15:220.
14. van Staa TP, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, Cassell J, Pirmohamed M, Torgerson D, Ronaldson S, Adamson J, Taweel A, Delaney B, Mahmood S, Baracaia S, Round T, Fox R, Hunter T, Gulliford M, Smeeth L. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. Health Technol Assess. 2014;18:1-146.
15. Chisholm, J., The Read clinical classification. BMJ. 1990;300:1092.
16. Chen YC, Wu JC, Haschler I, Majeed A, Chen TJ, Wetter T. Academic impact of a public electronic health database: bibliometric analysis of studies using the general practice research database. PLoS One. 2011;6:e21404.

17. Clinical Practice Research Datalink. CPRD Bibliography. Available at: http://www.cprd.com/bibliography/ [Accessed 12 February 2015]
18. Staa TP, Goldacre B, Gulliford M, Cassell J, Pirmohamed M, Taweel A, Delaney B, Smeeth L. Pragmatic randomised trials using routine electronic health records: putting them to the test. BMJ. 2012;344:e55.
19. Smeeth L, Cook C, Fombonne E, Heavey L, Rodrigues LC, Smith PG, Hall AJ. MMR vaccination and pervasive developmental disorders: a case-control study. Lancet. 2004;364:963-69.
20. Smeeth L, Thomas SL, Hall AJ, Hubbard R, Farrington P, Vallance P. Risk of myocardial infarction and stroke after acute infection or vaccination. N Engl J Med. 2004;351:2611-18.
21. Jick H, Zornberg GL, Jick SS, Seshadri S, Drachman DA. Statins and the risk of dementia. Lancet. 2000;356:1627-31.
22. Gelfand JM, Neimann AL, Shin DB, Wang X, Margolis DJ, Troxel AB. Risk of myocardial infarction in patients with psoriasis. JAMA. 2006;296:1735-41.
23. Van Staa TP, Leufkens HG, Abenhaim L, Zhang B, Cooper C. Use of oral corticosteroids and risk of fractures. J Bone Miner Res. 2000;15:993-1000.
24. Bhaskaran K, Douglas I, Forbes H, dos-Santos-Silva I, Leon DA, Smeeth L. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. Lancet. 2014;384:755-65.
25. Dommett RM, Redaniel T, Stevens MC, Martin RM, Hamilton W. Risk of childhood cancer with symptoms in primary care: a population-based case-control study. Br J Gen Pract. 2013;63:e22-29.
26. Douglas I, Evans S, Rawlins MD, Smeeth L, Tabrizi SJ, Wexler NS. Juvenile Huntington's disease: a population-based study using the General Practice Research Database. BMJ Open. 2013;3:e002085.
27. Crooks CJ, Card TR, West J. Excess long-term mortality following non-variceal upper gastrointestinal bleeding: a population-based cohort study. PLoS Med. 2013;10:e1001437.
28. Cotton SJ, Belcher J, Rose PK, Jaqadeesan S, Neal RD. The risk of a subsequent cancer diagnosis after herpes zoster infection: primary care database study. Br J Cancer. 2013;108:721-26.
29. Lalmohamed A, Bazelier MT, Van Staa TP, Uitdehaag BM, Leufkens HG, De Boer A, De Vries F. Causes of death in patients with multiple sclerosis and matched referent subjects: a population-based cohort study. Eur J Neurol. 2012;19:1007-14.
30. Office for National Statistics, 2011 Census - Population and Household Estimates for England and Wales, March 2011. Cardiff, UK: ONS, 2012.
31. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, vanStaa T, Grundy E, Smeeth L. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health (Oxf). 2013;34:684-92.
32. Bhaskaran K, Forbes HJ, Douglas I, Leon DA, Smeeth L. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). BMJ Open. 2013;3:e003389.
33. Campbell J, Dedman DJ, Eaton SC, Gallagher AM, Williams TJ. Is the GPRD GOLD population comparable to the UK population? Pharmacoepidemiol Drug Saf. 2013;22(Suppl 1):280.
34. Quality and Outcomes Framework. Available at: http://www.hscic.gov.uk/qof [6 August 2014]
35. van Staa TP, Dennison EM, Leufkens HG, Cooper C. Epidemiology of fractures in England and Wales. Bone. 2001;29:517-22.
36. Ryan R, Majeed A. Prevalence of treated hypertension in general practice in England and Wales, 1994 and 1998. Health Stat Q. 2002;16:14-18.

37. Ronquist G, Rodríguez LA, Ruigómez A, Johansson S, Wallander MA, Frithz G, Svärdsudd K. Association between captopril, other antihypertensive drugs and risk of prostate cancer. Prostate. 2004;58:50-56.
38. Meier CR, Napalkov PN, Wegmuller Y, Jefferson T, Jick H. Population-based study on incidence, risk factors, clinical complications and drug utilisation associated with influenza in the United Kingdom. Eur J Clin Microbiol Infect Dis. 2000;19:834-42.
39. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69:4-14.
40. Tate R, Kalra D, Boggon R, Beloff N, Puri S, Williams TJ. Data quality in European primary care research databases. Report of a workshop held in London September 2013. 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI. 2014;6864310:85-88.
41. Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. Pharmacoepidemiol Drug Saf. 2009;18:704-07.
42. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, Reeves D. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PLoS One. 2014;9:e99825.
43. CALIBER Research Portal. Available at: http://www.caliberresearch.org/portal/ [Accessed 23 January 2015]

Supporting information

Appendix S1: Glossary of terms/data definitions

Acceptable Patients

2.1

Patients are labelled as ‘acceptable’ for use in research by a process that identifies and excludes patients with non-continuous follow up or patients with poor data recording that raises suspicion as to the validity of the that patients record. Patient data is checked, for the following issues:

- An empty or invalid first registration date
- An empty or invalid current registration date
- Absence of a record for a year of birth
- A first registration date prior to their birth year
- A current registration date prior to their birth year
- A transferred out reason with no transferred out date
- A transferred out date with no transferred out reason
- A transferred out date prior to their first registration date
- A transferred out date prior to their current registration date
- A current registration date prior to their first registration date
- A gender other than Female/Male/Indeterminate
- An age of greater than 115 at end of follow up
- Recorded healthcare episodes in years prior to birth year
- All recorded healthcare episodes have empty or invalid event dates
- Registration status of temporary patients

If any of these conditions are true then the patient is labelled unacceptable, and is not recommended for use in research.

UTS date

The overall quality of data in practices is mediated by use of an ‘up to standard’ (UTS) date, which is deemed as the date at which data in the practice is considered to have continuous high quality data fit for use in research. This is mediated by an analysis on the total data in the practice, which is refreshed every time a new collection for a practice is processed into the database. It is based on two central concepts: assurance of continuity in data recording (gap analysis), and avoidance of use of data for which transferred out and dead patients have been removed (death recording).

Gap Analysis

To detect whether there is any meaningful gaps in the data it is necessary to look in more detail at single day gaps as well as longer gaps. A single day alone may reflect a situation where nothing was recorded that day at the practice, i.e. the practice was not open, such as on a bank holiday. A longer gap may reflect a situation where the practice did not offer a service and patients may have been treated elsewhere. If a meaningful gap is found, the earliest date after which there is no significant gap is identified.

Death Recording

It is expected that a standard number of deaths will be recorded at a practice over time. Assessment of gaps in death recording is performed taking the size of the practice into account. A safety margin is built in to account for both geographical and seasonal variation in death rates. If a meaningful gap is found, the earliest date after which there is no significant gap is identified.

The UTS date is set to the latest of these dates for each practice. The CPRD recommend that analyses are performed on data following the practice UTS date.

Table S1: Data sources linked to CPRD primary care records.

Linkage	HES inpatient	HES outpatient	MINAP	ONS	National Cancer Data Repository (NCDR)	Deprivation
Type of resource	Hospitalisation data for inpatients	Outpatient data	Acute coronary syndrome disease registry	Date and cause of death register	Cancer registry	Socioeconomic status
Who is included?	Patients with hospitalisations for any cause	Patients presenting in outpatients	Patients hospitalised with acute coronary syndrome in NHS hospitals	People who die in England and Wales	Tumour level records submitted to Office of National Statistics (ONS) by the England Cancer Registries	Lower super output area levels
Geographic regions covered by linkage	England	England	England and Wales	England and Wales	England	England, Wales, Scotland and Northern Ireland
Period of linkage	1997 onwards	2003 onwards	2003 onwards	Cause of death recorded since 1841, but since 2001 using ICD-10 codes,	1990 onwards	Calculated at different times for each country
Examples of data available in linked dataset	Diagnoses, procedures	Diagnoses	Details regarding ACS diagnosis and management	Date and cause of death, including underlying and secondary causes.	Cancer type and date of diagnosis	Townsend Scores (material deprivation), Index of Multiple Deprivation (IMD)

Table S1 (cont.): Data sources linked to CPRD primary care records.

Linkage	HES inpatient	HES outpatient	MINAP	ONS	National Cancer Data Repository (NCDR)	Deprivation
How are data coded	ICD-10 and OPCS-4	ICD-10	In 120 fields with multiple response categories as defined by the MINAP steering group.	ICD-10	ICD-10	Townsend Scores, IMD - in quintiles
Permissions required	ISAC	ISAC	ISAC, MINAP academic group	ISAC	ISAC, PHE	ISAC
Additional cost for data?	Yes	Yes	Yes	No	Yes	No

Note: These data are available for 75% of English practices (58% of all practices) in CPRD

ACS: Acute Coronary Syndrome; CPRD: Clinical Practice Research Datalink; HES: Hospital Episode Statistics; ICD-10: International Classification of Disease, version 10; ISAC: Independent Scientific Advisory Committee; MINAP: Myocardial Ischaemia National Audit Project; ONS: Office for National Statistics; OPCS-4: Office of Population, Censuses and Surveys Classification of Surgical operations and procedures, version 4 codes; PHE: Public Health England.

Table S2: Data files supplied by the Clinical Practice Research Datalink (CPRD)

File type	What it holds	Example of contents
Patient	Demographic and registration status of patients	Patient identifier, month and year of birth, registration status, death date, transfer out date
Practice	Practice administrative data	Practice identifier, geographical region, date practice became 'Up to standard', last data collection date
Staff	Information about the staff members entering data	Staff identifier, gender, role
Consultation	Administrative information about the consultation	Date of clinical event, date of data entry, type of consultation, staff identifier and duration of consultation
Clinical	Clinical data regarding medical history	Date of clinical event, date of data entry, the CPRD medical code for the chosen Read code, additional details identifier*, entity type
Additional Clinical Details (ACD)	Specific data about a clinical event	Type of information held, called an 'entity', specific clinical details relating to that entity
Referral	Details on referrals to secondary care or specialists	The CPRD medical code for the chosen Read code, method of referral, referral specialty, urgency of referral
Immunisation	Data associated with immunisations	Reason for immunisation, type, stage, status and the compound used
Test	Test results	Type of test, result, normal range of result, unit of measure
Therapy	Information about therapies including medications and appliances	The CPRD product code for the medication, British National Formulary code, quantity of product, dose, pack size, number of days prescribed

*Allows a link to be made between a Read code in the 'clinical file' to additional details held in the 'additional clinical details' file.

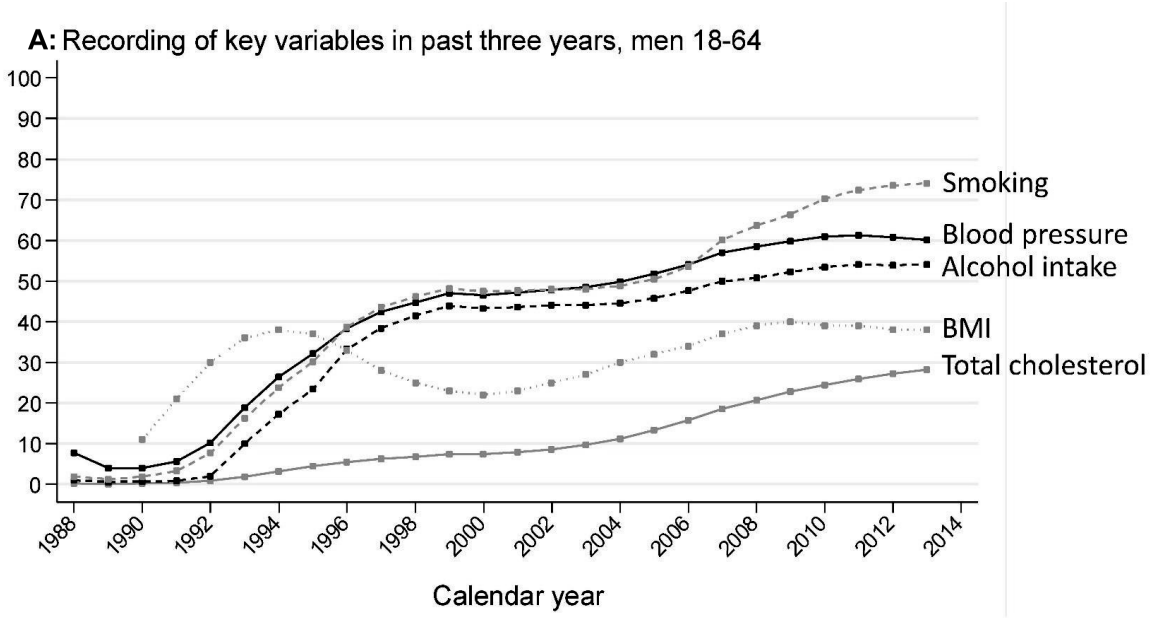


Figure S1: Recording of key lifestyle and demographic variables in the past three years of patient follow-up, by calendar year (A: for men aged 18-64; B: for men aged 65 or over).

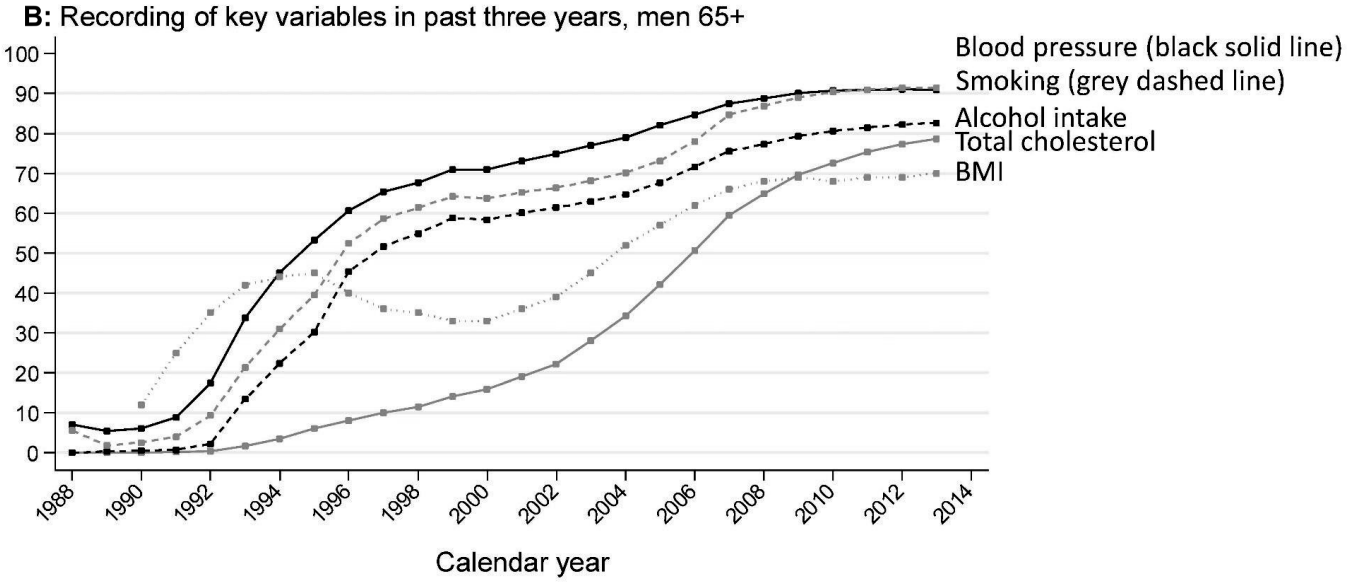


Figure S1 (cont.): Recording of key lifestyle and demographic variables in the past three years of patient follow-up, by calendar year (A: for men aged 18-64; B: for men aged 65 or over).

Note: These data are based on a one million patient sample of primary care data from the CPRD. All patients are acceptable. Records outside of UTS were included. Denominators are mid-year registered populations. Total cholesterol has poorer completeness as this would not be routinely recorded and would require a clinical indication.
BMI: body mass index; CPRD: Clinical Practice Research Datalink; UTS: Period of up-to-standard data recording.

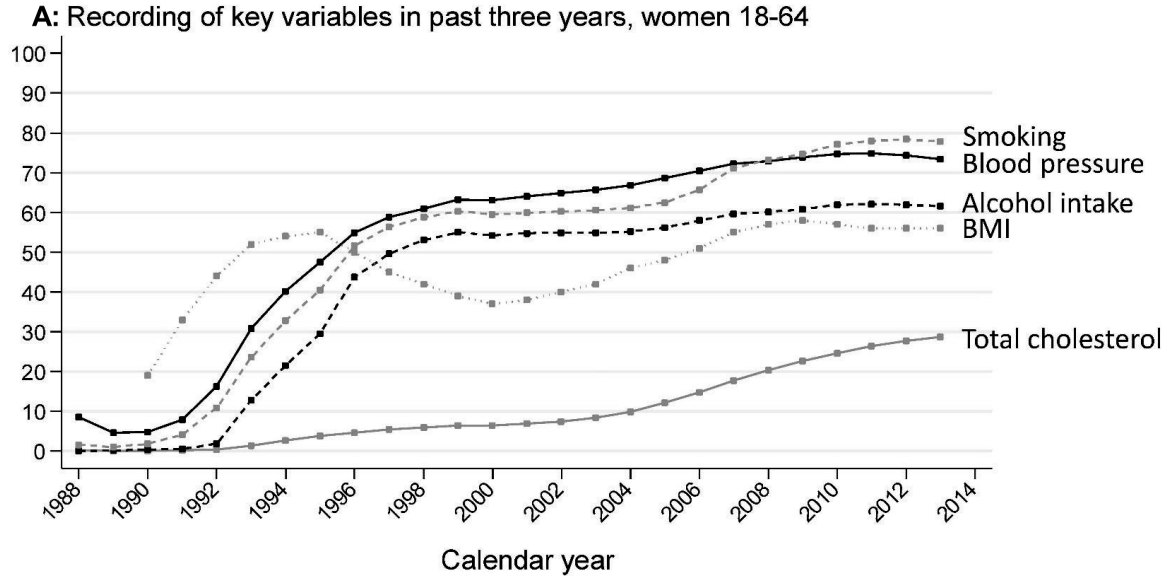


Figure S2: Recording of key lifestyle and demographic variables recorded in the past three years of patient follow-up, by calendar year
(A: for women aged 18-64; B: for women aged 65 or over).

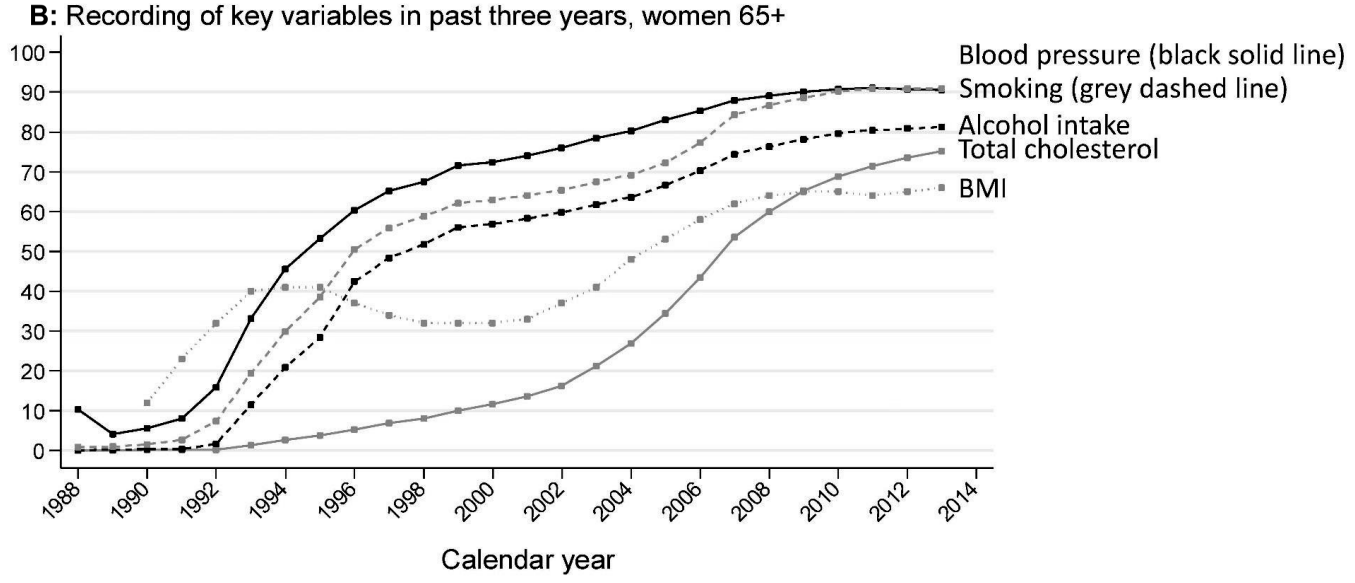


Figure S2 (cont.): Recording of key lifestyle and demographic variables recorded in the past three years of patient follow-up, by calendar year
(A: for women aged 18-64; B: for women aged 65 or over).

Note: These data are based on a one million patient sample of primary care data from the CPRD. All patients are acceptable. Records outside of UTS were included. Denominators are mid-year registered populations. Total cholesterol has poorer completeness as this would not be routinely recorded and would require a clinical indication.
BMI: body mass index; CPRD: Clinical Practice Research Datalink; UTS: Period of up-to-standard data recording.

Chapter 2.2

Comparability of the age and sex distribution of the UK Clinical Practice Research Datalink and the total Dutch population

Roy G.P.J. de Jong, Arlene M. Gallagher, Emily Herrett,
Ad A.M. Masclee, Maryska L.G. Janssen-Heijnen, Frank de Vries

Note: RdJ and AG are joint first author.

Pharmacoepidemiol Drug Saf. 2016;25(12):1460-4

Abstract

Purpose: The UK Clinical Practice Research Datalink (CPRD) is increasingly being used by Dutch researchers in epidemiology, and pharmaco-epidemiology. It is however unclear if CPRD is representative of the Dutch population and whether study results would apply to the Dutch population. Therefore, as first step, our objective was to compare the age and sex distribution of the CPRD with the total Dutch population.

Methods: As a measure of representativeness, the age and sex distribution of the CPRD were visually and numerically compared with Dutch census data from the StatLine database of the Dutch National Bureau of Statistics in 2011.

Results: The age distribution of men and women in the CPRD population was comparable to the Dutch male and female population. Differences of more than 10% occurred only occurred in older age categories (75+ in men, and 80+ in women).

Conclusions: Results from observational studies that have used CPRD data are applicable to the Dutch population, and a useful resource for decision making in the Netherlands. Nevertheless, differences in drug exposure likelihood between countries should be kept in mind, as these could still cause variations in the actual population studied, thereby decreasing its generalisability.

Keywords: database, epidemiology, pharmacoepidemiology, population-based, representativeness

Introduction

The United Kingdom (UK) Clinical Practice Research Datalink (CPRD) is one of the world's largest primary care databases, and is frequently used for post-authorisation safety studies, pharmaco-epidemiology, and disease epidemiology [1-5]. Examples include the evaluation of side effects of dopamine agonists [2] or diabetes drugs [3, 4], and the epidemiology of fractures [5].

From a global perspective, the healthcare system in the Netherlands, a small country not far from the UK, is largely comparable to that of the UK's National Health Service (NHS); i.e. everyone has equal access to medical care regardless of income or socioeconomic status. In both countries, the general practitioner (GP) is the gatekeeper of the public healthcare system, meaning patients cannot refer themselves to secondary or tertiary care without the GP's approval. These conditions are key for conducting population-based pharmaco-epidemiological studies.

Although the Dutch public healthcare system has excellent conditions for establishing a large primary care database for pharmaco-epidemiological research that is comparable to the CPRD in terms of sample size, this has not yet occurred. Smaller primary care databases do exist (e.g. Netherlands Information Network of General Practice (LINH), Integrated Primary Care Information (IPCI), PHARMO Database Network); however these are generally restricted by a limited set of medical codes (approximately 1000 different "International Classification of Primary Care" codes versus over 100,000 Read codes in the CPRD), few validation studies, considerable smaller sample size (e.g. 350.000 in LINH and 1.5 million in IPCI versus over 11 million in CPRD in 2011), and limited access to routinely collected lifestyle data, such as tobacco use, alcohol consumption, and socioeconomic status [6]. Furthermore, claims that these data are representative for the total Dutch population are seldom supported by published figures [7, 8]. For these reasons, an increasing number of CPRD studies are being conducted by researchers from the Netherlands and are financially supported by Dutch universities and funding agencies such as ZonMw and NWO. A recent study showed that the CPRD is representative of the total UK population with respect to age and sex and covers 6.9% of the UK population [1, 9]. A wide range of diagnoses in the CPRD have been validated in a number of studies and data quality are further enhanced by NHS annual reward and incentive programme that details GP practice achievement results, the Quality and Outcomes Framework (QOF). The QOF awards GPs for regular recording of detailed data on a wide range of diseases. As a result, the CPRD contains millions of recordings for measurements such as blood pressure, cholesterol values and lung function. In addition, the strength of CPRD's data partially explain why officials of the Food and Drug Administration in the United States perform studies using the CPRD database for drug safety monitoring and regulatory decision making [10-12].

Although there are many similarities between healthcare systems of the UK and the Netherlands, it is unclear whether results from CPRD studies would apply to the Dutch population. Therefore, our objective is to compare the age and sex distribution of the CPRD to the total Dutch population.

Methods

Using the same sample of data from the previously published CPRD data resource profile (Chapter 2.1), the age and sex distribution of the CPRD primary care data on 27 March 2011 were visually and numerically compared with UK and Dutch census data in 2011 [1]. The CPRD (formerly known as Value Added Medical Products (VAMP), and later General Practice Research Database (GPRD) [13, 14]) harnesses data from UK's general practices and produces a primary care dataset since 1987. Through the years, it has become one of the largest databases of longitudinal medical records from primary care in the world, with coverage of over 11.3 million patients from 674 practices. To date, 4.4 million active (alive, currently registered) patients meet quality criteria (approximately 6.9% of the UK population), who are broadly representative of the UK general population in terms of age, sex and ethnicity. For this study, visual comparison was performed by inspecting the overlap between the respective lines in a graph (Figure 1). An additional comparison calculated the differences between proportional distributions of 5-year age groups of CPRD data versus Dutch census data in 2011. We are not aware of any objective methods to define representativeness of patients in a research database compared to a country's total population. We therefore described the absolute and proportional differences between the age and sex distributions of CPRD and Dutch Census data in order to leave this to the reader, and made a subjective decision to consider an age-sex specific difference of <10% representative (Table 1). Numbers for computing sex-stratified age categories of the total Dutch population in March 2011 were obtained from the StatLine database of the Dutch National Bureau of Statistics (www.cbs.nl).

Results

In general, the age distribution of men and women in the CPRD population was comparable to the Dutch male and female population (Figure 1). Overall, the percentage men and women in the CPRD in 2011 was the same as in the Dutch population (49.5% men, 50.5% women). The additional comparison based on calculating the differences between proportional distributions of 5-year age groups, showed that differences of more than 10% occurred only in older age categories, starting from 75+ in men, and 80+ in women (Table 1).

Table 1: Differences between proportional age distributions of CPRD and the total Dutch population

Age group	Total				Men				Women			
	CPRD	NL census	Δ*	Δ%†	CPRD	NL census	Δ*	Δ%†	CPRD	NL census	Δ*	Δ%†
0-4	4.99	5.54	-0.55	-9.94	5.15	5.72	-0.57	-9.98	4.82	5.35	-0.53	-9.90
5-9	5.61	5.89	-0.28	-4.75	5.80	6.09	-0.29	-4.74	5.43	5.70	-0.27	-4.77
10-14	5.68	6.00	-0.32	-5.38	5.87	6.20	-0.33	-5.32	5.49	5.80	-0.31	-5.43
15-19	5.99	6.03	-0.04	-0.65	6.18	6.23	-0.05	-0.80	5.80	5.83	-0.03	-0.50
20-24	5.83	6.23	-0.40	-6.44	5.98	6.35	-0.37	-5.89	5.68	6.10	-0.42	-7.00
25-29	6.32	6.02	0.30	4.96	6.34	6.13	0.21	3.41	6.31	5.92	0.39	6.54
30-34	6.51	6.03	0.47	7.83	6.57	6.11	0.46	7.65	6.44	5.96	0.48	8.01
35-39	6.69	6.67	0.01	0.20	6.84	6.74	0.10	1.62	6.53	6.61	-0.08	-1.22
40-44	7.59	7.77	-0.17	-2.22	7.80	7.91	-0.11	-1.42	7.39	7.62	-0.23	-3.03
45-49	7.77	7.79	-0.02	-0.31	7.97	7.95	0.02	0.22	7.58	7.64	-0.06	-0.86
50-54	6.95	7.19	-0.24	-3.41	7.12	7.30	-0.18	-2.44	6.78	7.09	-0.31	-4.38
55-59	6.06	6.56	-0.49	-7.54	6.14	6.64	-0.50	-7.62	5.99	6.47	-0.48	-7.46
60-64	6.15	6.64	-0.48	-7.24	6.16	6.72	-0.56	-8.30	6.15	6.55	-0.40	-6.18
65-69	5.30	4.79	0.51	10.64	5.23	4.79	0.44	9.18	5.37	4.79	0.58	12.08
70-74	4.03	3.83	0.20	5.27	3.87	3.67	0.20	5.31	4.19	3.98	0.21	5.23
75-79	3.36	3.00	0.36	11.90	3.09	2.66	0.43	15.98	3.62	3.33	0.29	8.71
80-84	2.56	2.17	0.39	18.00	2.15	1.70	0.45	26.28	2.97	2.64	0.33	12.80
85-89	1.64	1.28	0.36	28.27	1.20	0.82	0.38	47.05	2.07	1.73	0.34	19.62
90+	0.98	0.57	0.41	70.52	0.55	0.27	0.28	102.20	1.40	0.87	0.53	60.92

Δ Difference.

* Absolute difference.

† Percentage difference between CPRD versus NL Census data (CPRD minus NL Census, divided by NL Census), with >10% regarded as different (BOLD).

CPRD: Clinical Practice Research Datalink; NL: Netherlands.

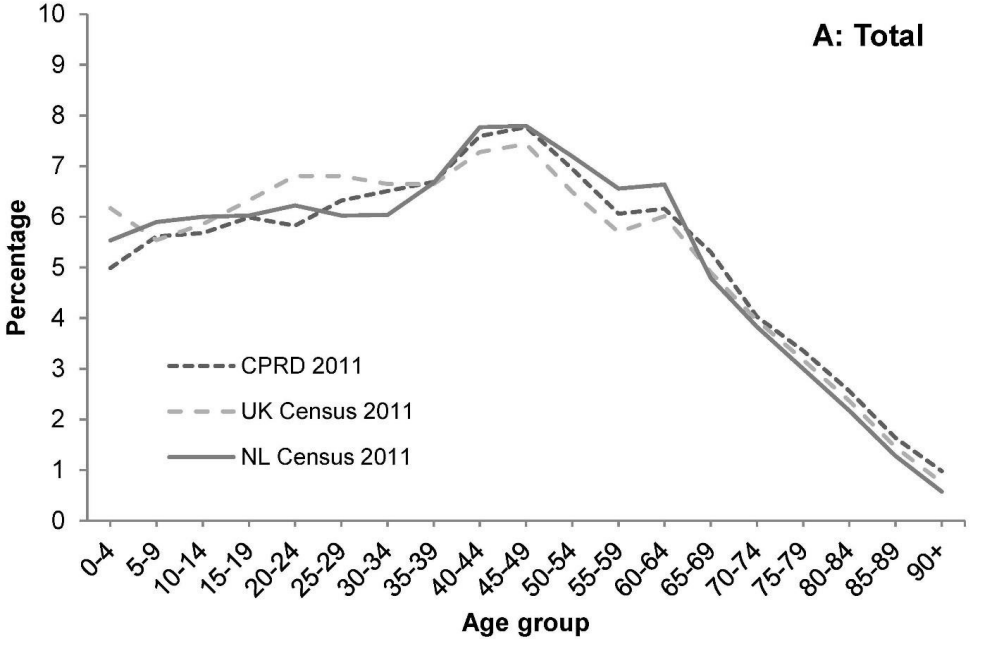


Figure 1: Age distribution of the CPRD primary care data on 27 March 2011 compared with UK and NL Census data from 2011, in both men and women (A), men only (B) and women only (C).

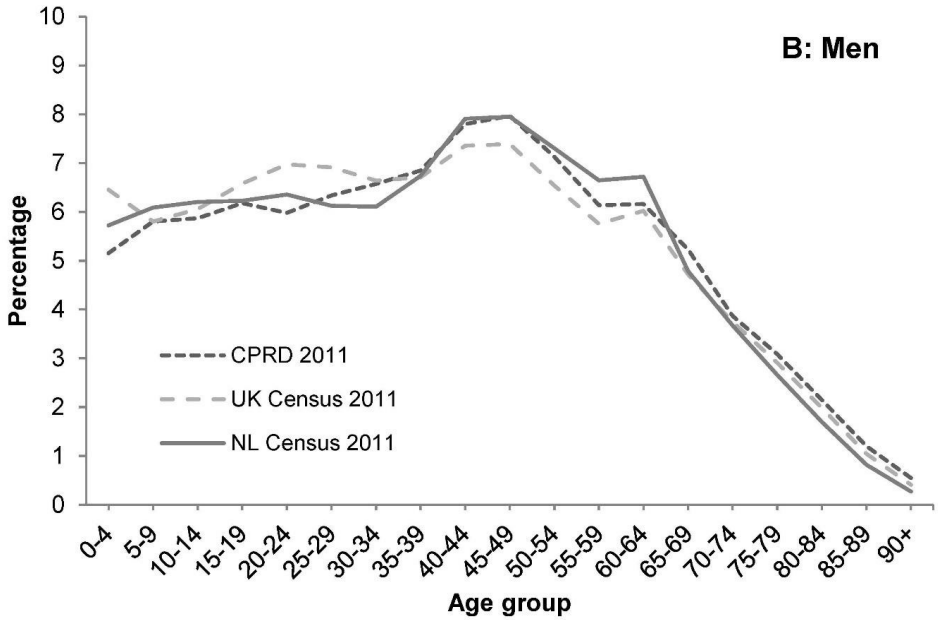


Figure 1 (cont.): Age distribution of the CPRD primary care data on 27 March 2011 compared with UK and NL Census data from 2011, in both men and women (A), men only (B) and women only (C).

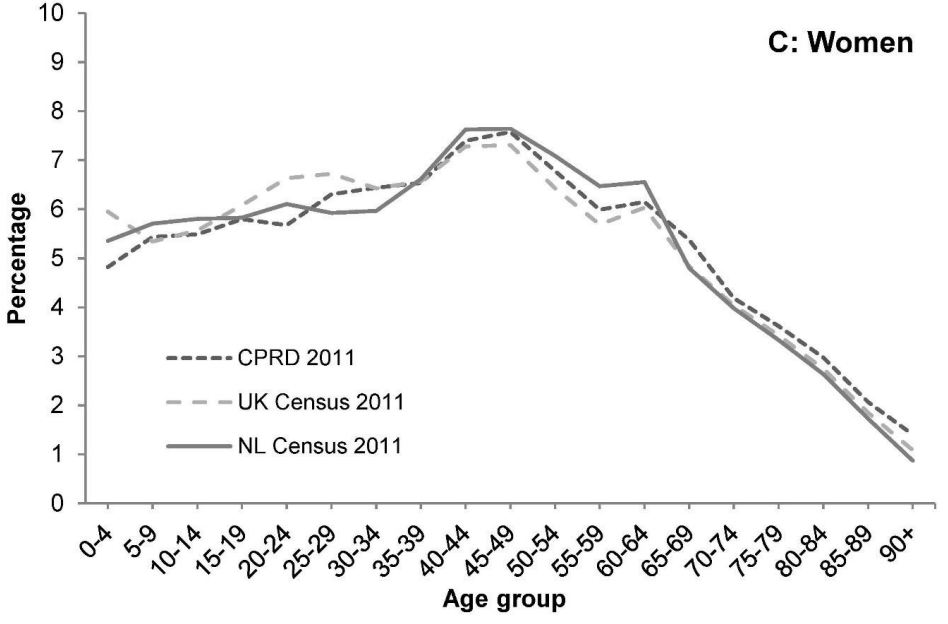


Figure 1 (cont.): Age distribution of the CPRD primary care data on 27 March 2011 compared with UK and NL Census data from 2011, in both men and women (A), men only (B) and women only (C).

Note: These data are based on a one-million patient sample of CPRD. Adapted from Herrett et al. Int. J Epidemiol 2015.

Discussion

In this study, we showed that the age and sex distribution the CPRD was visually and numerically comparable to that of the total Dutch population in 2011.

To our knowledge, this is the first study to directly compare the age and sex distribution of the CPRD to the total Dutch population. Apart from LINH, a national data network compiling electronic medical records (EMR) of 92 primary care practices with 211 GPs and over 350,000 patients, we are not aware of any Dutch EMR database with data from GPs, which has published data on its representativeness according to the total Dutch population. The IPCI database is a longitudinal primary care database maintained by the department of Medical Informatics of the Erasmus Medical Centre in Rotterdam. In published papers of studies using the IPCI database it is frequently stated that the database is comparable to the total Dutch population in terms of age and sex [15, 16]. However, we could not verify this claim in a (peer-reviewed) publication. In a report of the Netherlands institute for health services research (NIVEL), which maintains the LINH database, it is shown that LINH is generally comparable to the total Dutch population, with a slight underrepresentation of women of 75 years and older [8]. However, age- and sex-stratified proportions were calculated as compared to the total LINH population and not as compared to the total LINH population stratified by sex, making direct comparison with our results difficult.

Our calculations of the differences between proportional distributions of 5-year age groups show that the CPRD is comparable to the Dutch population in terms of age and sex up to age 75 in men and age 80 in women. However, in Figure 1 it can be clearly seen that the age distribution of CPRD and Dutch census data almost overlap in these higher age groups. The large differences for higher age groups seen in Table 1 may be a spurious finding, since the calculations were based on very small proportions of the total population. Of note, some difference is also seen between CPRD and the UK population in Figure 1. Men and to a lesser extent women aged between 20 and 35 years of age are underrepresented in CPRD, which has been attributed to the fact that these individuals probably do not register with a GP [9].

Several strengths have to be noted for this study. First, UK CPRD data were compared to data from the StatLine database of the Dutch National Bureau of Statistics, a reliable source of population-based information, regulated by national and European codes and laws [17]. Second, by calculating the differences between proportional distributions in all age groups, we demonstrated that the distributions were not only visually, but also numerically comparable. Last, by showing the overall representativeness of the CPRD database of another population besides the UK

population, CPRD may be used as rich data source for healthcare policymakers outside the UK.

There are also several limitations to this comparison. First, there are no objective methods to define representativeness of patients in a research database compared to a country's total population. To overcome this, we gave the reader insight into the various ways of comparing these data. Furthermore, the CPRD population was compared to the total Dutch population in terms of age and sex only. Therefore, we cannot rule out differences in for instance ethnicity, socioeconomic class, or lifestyle, which may in turn impact disease prevalence, exposure to important risk factors, or the degree of health care seeking behaviour. Based on the report of the OECD health indicators, the UK population has higher rates of tobacco and alcohol consumption, and especially obesity among adults, compared to the Dutch population [18]. Although the results of our study imply that the UK CPRD may also be representative of the total Dutch population, information on a specific population of drug-users is ultimately necessary to know whether a study's results are transferrable to one's own region. As of yet, we have not looked into the comparability of various subpopulations in CPRD and The Netherlands. In addition, relative risks of disease outcomes found in CPRD could be extrapolated to the Dutch population, incidence rates or absolute risks cannot.

In conclusion, this study showed that the age- and sex distribution of CPRD were generally comparable to that of the total Dutch population. Results from observational studies that have used CPRD data are applicable to the total Dutch population (similar to how relative risks from randomised clinical trials apply to their demarcated population), and a useful resource for decision making in the Netherlands. Nevertheless, differences in drug exposure likelihood between countries should be kept in mind, as these could still cause variations in the actual population studied, thereby decreasing its generalisability. In addition, the results of this study may encourage scientists from other countries with similar healthcare systems to perform comparable studies of CPRD representativeness.

Key points

- The age and sex distribution of the UK CPRD were comparable to the total Dutch population.
- Results from observational studies that have used CPRD data are applicable to the Dutch population.

References

1. Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36.
2. Schade R, Andersohn F, Suissa S, *et al*. Dopamine agonists and the risk of cardiac-valve regurgitation. N Engl J Med. 2007;356(1):29-38.
3. Knapen LM, van Dalem J, Keulemans YC, *et al*. Use of Incretin Agents and Risk of Pancreatic Cancer: A Population-Based Cohort Study. Diabetes Obes Metab. 2015;18(3):258-65.
4. Lo Re III V, Haynes K, Ming EE, *et al*. Safety of saxagliptin: rationale for and design of a series of postmarketing observational studies. Pharmacoepidemiol Drug Saf. 2012;21(11):1202-15.
5. van Staa TP, Dennison EM, Leufkens HG, Cooper C. Epidemiology of fractures in England and Wales. Bone. 2001;29(6):517-22.
6. Requena G, Huerta C, Gardarsdottir H, *et al*. Hip/femur fractures associated with the use of benzodiazepines (anxiolytics, hypnotics and related drugs): a methodological approach to assess consistencies across databases from the PROTECT-EU project. Pharmacoepidemiol Drug Saf. 2015;25(S1):66-78.
7. Abbing-Karahagopian V, Kurz X, de Vries F, *et al*. Bridging differences in outcomes of pharmacoepidemiological studies: design and first results of the PROTECT project. Curr Clin Pharmacol. 2014;9(2):130-8.
8. NIVEL. Representativiteit van LINH. 2010.
9. Campbell J, Dedman DJ, Eaton SC, Gallagher AM, Williams TJ. Is the GPRD GOLD population comparable to the UK population? Pharmacoepidemiol Drug Saf. 2013;22(Suppl 1):280.
10. Zhou EH, Gelperin K, Levenson MS, *et al*. Risk of acute myocardial infarction, stroke, or death in patients initiating olmesartan or other angiotensin receptor blockers - a cohort study using the Clinical Practice Research Datalink. Pharmacoepidemiol Drug Saf. 2014;23(4):340-7.
11. Hammad TA, Graham DJ, Staffa JA, *et al*. Onset of acute myocardial infarction after use of non-steroidal anti-inflammatory drugs. Pharmacoepidemiol Drug Saf. 2008;17(4):315-21.
12. US Food and Drug Administration (FDA). FDA Drug Safety Communication: Possible fracture risk with high dose, long-term use of proton pump inhibitors. 2010.
13. Kousoulis AA, Rafi I, de Lusignan S. The CPRD and the RCGP: building on research success by enhancing benefits for patients and practices. Br J Gen Pract. 2015;65(631):54-5.
14. Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. Ther Adv Drug Saf. 2012;3(2):89-99.
15. Noordam R, Aarts N, Verhamme KM, *et al*. Prescription and indication trends of antidepressant drugs in the Netherlands between 1996 and 2012: a dynamic population-based study. Eur J Clin Pharmacol. 2015;71(3):369-75.
16. Masclee GM, Coloma PM, Spaander MC, *et al*. NSAIDs, statins, low-dose aspirin and PPIs, and the risk of oesophageal adenocarcinoma among patients with Barrett's oesophagus: a population-based case-control study. BMJ Open. 2015;5(1):e006640.
17. Centraal Bureau voor de Statistiek. CBS Kwaliteitsverklaring. 2014.
18. OECD. Health at a Glance 2015: OECD Indicators. Paris: OECD Publishing, 2015.

Chapter 3:

Studying mortality using CPRD

Chapter 3.1

Risks of stroke and mortality associated with suboptimal anticoagulation in atrial fibrillation patients

Arlene M. Gallagher, Efrosini Setakis, Jonathan M. Plumb,
Andreas Clemens, Tjeerd-Pieter van Staa

Thromb Haemost. 2011;106(5):968-77

Summary

Atrial fibrillation (AF) carries an increased risk of ischaemic stroke, and oral anticoagulation with warfarin can reduce this risk. The objective of this study was to evaluate the association between time in therapeutic International Normalised Ratio (INR) range when receiving warfarin and the risk of stroke and mortality. The study cohort included AF patients aged 40 years and older included in the UK General Practice Research Database. For patients treated with warfarin we computed the percentage of follow-up time spent within therapeutic range. Cox regression was used to assess the association between INR and outcomes while controlling for patient demographics, health status and concomitant medication. The study population included 27,458 warfarin-treated (with at least 3 INR measurements) and 10,449 patients not treated with antithrombotic therapy. Overall the warfarin users spent 63% of their time within therapeutic range. This percentage did not vary substantially by age, sex and CHA_2DS_2-VASc score. Patients who spent at least 70% of time within therapeutic range had a 79% reduced risk of stroke compared to patients with ≤ 30% of time in range (adjusted relative rate of 0.21; 95% confidence interval 0.18-0.25). Mortality rates were also significantly lower with at least 70% of time spent within therapeutic range. In conclusion, good anticoagulation control was associated with a reduction in the risk of stroke.

Keywords: anticoagulation, atrial fibrillation, stroke, treatment, warfarin

Introduction

Atrial fibrillation (AF) is the most common sustained disorder of cardiac rhythm. The condition carries an increased risk of arterial thromboembolism and ischaemic stroke due to embolisation of thrombi that form within the left atrium of the heart [1]. Long-term anticoagulation therapy of patients with non-valvular AF can reduce the risk of stroke by two-thirds, as reported in a meta-analysis of clinical trials [2]. However, a recent large epidemiological study found that the risk of stroke was only reduced by 38% in current warfarin users compared to past users. Use of aspirin did not appear to have major beneficial effects [3]. The effectiveness of warfarin is dependent on the quality of anticoagulation, with optimal risk reduction and appropriate safety occurring within the recommended International Normalised Ratio (INR) range of 2.0 to 3.0. The risks of ischaemic and intracranial haemorrhage are significantly higher for INR values below 2.0 and above 3.5, respectively [3-5].

A systematic review has reported that patients who had received anticoagulation therapy spend a significant percentage of their time with an INR outside therapeutic range [6]. Reduced anticoagulation control in warfarin users has been associated with an increased risk of stroke and mortality [7-9]. The study size of these studies was limited and these studies compared warfarin users to patients untreated with warfarin. It has been reported that warfarin is preferentially prescribed to patients at lower rate of stroke - contrary to guidelines - and who are less frail [3,10], possibly confounding statistical comparisons. The objective of this study was to evaluate the association between time in therapeutic range and risk of stroke and all-cause mortality. The main analyses consisted of comparisons of warfarin users with different percentages of time spent within therapeutic range.

Material and methods

Source of data

Data for this study were obtained from the General Practice Research Database (GPRD). GPRD collates the computerised medical records of general practitioners (GPs). GPs play a key role in the UK healthcare system, as they are responsible for primary healthcare and specialist referrals. Patients are semi-permanently affiliated with a practice that centralises the medical information from the GPs, specialist referrals, and hospitalisations. The data recorded in the GPRD include demographic information, prescription details, clinical events, preventive care provided, specialist referrals, hospital admissions, and major outcomes [11]. The protocol of this study was approved by GPRD's Independent Scientific Advisory Committee.

Study population

The study population consisted of men and women aged 40 years or older registered with a GPRD practice and with a diagnosis of AF (including paroxysmal and persistent AF). Patients were excluded if they had a history of heart valve problems and/or valve replacement surgery. The study population was categorised according to use of warfarin. Warfarin users were study patients with at least one International Normalised Ratio (INR) measurement in their medical history. The start of follow-up (i.e., index date) was defined as their first INR measurement in their GPRD follow-up period. AF patients not treated with antithrombotic therapy were defined as study patients with no record ever of INR measurements or prescribing of warfarin, aspirin, clopidogrel, or dipyridamole (dipyridamole is often administered in a fixed-dose combination with aspirin). The index date for AF patients not treated with antithrombotic therapy was defined as the date of their first AF diagnosis in the GPRD follow-up period.

International Normalised Ratio (INR)

Warfarin use requires frequent blood tests to monitor the level of anticoagulation which is measured by the INR. INR values of warfarin users typically are recorded in GPRD if the laboratory sends the results electronically to the general practice or the practice holds in-house anticoagulation clinics. We computed the percentage of time spent within therapeutic INR range of 2-3 [12]. This method assumes that INR values vary linearly between two measurements. Using this relationship, the date at which a patient achieved or departed from an INR level spent within therapeutic range was computed. This was then used to calculate the percentage of their follow-up time spent within therapeutic range [12].

Stroke outcomes and CHADS2 and CHA2DS2-VASc risk scores

The study population was followed for the occurrence of stroke and transient ischaemic attack (TIA) from index date to six months after the last INR measurement, the end of data collection, patient's death or transfer out of the practice, whichever date came first. The analysis evaluated the association between time in therapeutic range when receiving warfarin and the risk of stroke and mortality. The main analysis included all types of stroke (embolic or haemorrhagic). The reason for this was that most GPRD records of stroke did not contain details of the type of stroke (the GPs tend to record the occurrence of stroke without specifying the aetiology).

The baseline $CHADS_2$ and CHA_2DS_2-VASc scores for stroke risk were estimated. The $CHADS_2$ risk score is based on a point system where two points are assigned to a history of stroke or transient ischaemic attack and one point each is assigned for, congestive heart failure, history of hypertension, age of 75 years or older, and diabetes mellitus [13]. The CHA_2DS_2-VASc risk score denotes Cardiac failure or dysfunction, Hypertension, Age ≥75[Doubled], Diabetes, Stroke[Doubled] – Vascular disease, Age 65-74 and Sex category [Female]), whereby 2 points are assigned for a history of stroke or age ≥75; and 1 point each for age 65-74 years, a history of hypertension, diabetes, cardiac failure and vascular disease. A CHA_2DS_2-VASc score of 0 corresponds to 'low risk', a score of 1 to 'intermediate risk' and ≥2 to 'high risk' [14].

Statistical analyses

A descriptive analysis of the distribution of INR measurements was conducted. Cox proportional hazards regression models were used to estimate the relative rates (RRs) for the associations between risk over time of stroke and mortality and time spent within therapeutic range [12]. In this analysis, patients were followed from the date of the third INR measurement. The results were adjusted for age and sex and for $CHADS_2$ risk score of stroke (without age), smoking history, body mass index, socioeconomic status of practice, alcohol use, duration of AF at baseline and prescribing in the six months before of other antithrombotics (aspirin, aspirin + dipyridamole, clopidogrel), anti-arrhythmics, heart rate limiting drugs and medications that can interact with warfarin.

Results

The study population included 37,907 AF patients; 27,458 warfarin users and 10,449 patients not treated with antithrombotic therapy. The median duration of follow-up was 1.7 years for warfarin users and 1.5 years for patients not treated with antithrombotic therapy (Table 1). The calendar year of the index dates ranged from 1991 to 2007. The majority of warfarin users (89.9%) was considered to be at high risk according to the CHA_2DS_2-VASc score (score ≥2). Patients using warfarin had on average a higher CHA_2DS_2-VASc score compared to patients not treated with antithrombotic therapy (including aspirin, aspirin + dipyridamole, clopidogrel). Prescribing of anti-arrhythmics and heart rate-limiting drugs prior to the index date was more frequent in warfarin users compared to patients not treated with antithrombotic therapy.

Figure 1 shows the percentage of patients with INR measurements outside therapeutic range stratified by the number of measurements. The median number of INR measurements was 10 and the mean 23. A considerable number of INR measurements were outside therapeutic range. At the first INR measurement, approximately 40% was spent within therapeutic range. At the tenth INR measurement, this figure was approximately 60%. Over the total duration of warfarin therapy, 63.1% of the time was spent within therapeutic range (Table 2). This percentage did not vary substantially by age and sex. Time spent within therapeutic range was 63.1% in patients with a CHA_2DS_2-VASc ≥2.

Tables 3 and 4 show the association between rate of stroke (including and excluding TIA, respectively) and the percentage of time spent within therapeutic range. A total of 1459 strokes (including TIAs) occurred in the warfarin users and 898 in the AF patients not treated with antithrombotic therapy. Patients who spent at least 70% of time within range had the largest reduction in the risk of stroke (79%) compared to patients with ≤ 30% of time in range (Table 3). A similar risk reduction was found in patients with a CHA_2DS_2-VASc ≥2 (78%). Compared to no therapy, time spent in therapeutic range of >61% showed a significant risk reduction for stroke. The risk of mortality was 81% lower in warfarin users with the best anticoagulation control compared to those with the worst control (Table 5).

Figure 2 shows the Kaplan-Meier curves for time to stroke in the AF patients not treated with antithrombotic therapy and in warfarin users stratified by percentage of time spent within therapeutic range. Warfarin users with ≥70% of time spent within therapeutic range had the lowest risk of stroke while those with <30% and 31-40% in range had the highest risks of stroke.

Sensitivity analyses showed similar results when restricting the data to the year 2000 or later (adjusted RR of 0.22 [95% confidence interval (CI) 0.17 - 0.28) with ≥70%, 0.23 [95% CI 0.16 - 0.29] with 61-70%, 0.35 [95% CI 0.25 - 0.45] with 51-60%, 0.49 [95% CI 0.34 - 0.64] with 41-50% and 0.60 [95% CI 0.39 - 0.81] with 31-40% compared to ≤ 30%).

Table 1: Baseline characteristics of the warfarin users and AF patients not treated with antithrombotic therapy

		Total (N=37,907)	**Warfarin users (N=27,458)**	**AF patients not treated with antithrombotic therapy (N=10,449)**	**Crude OR (95% CI)**
Follow-up (days)	Median	476	458	537	
	Percentiles 1, 99	2, 3742	17, 3233	1, 4716	
Age (years)	40-59	4317 (11.4%)	2603 (9.5%)	1714 (16.4%)	Reference
	60-69	7239 (19.1%)	5986 (21.8%)	1253 (12.0%)	0.32 (0.29 - 0.34)
	70-79	13572 (35.8%)	10982 (40.0%)	2590 (24.8%)	0.36 (0.34 - 0.38)
	80+	12779 (33.7%)	7887 (28.7%)	4892 (46.8%)	0.94 (0.89 - 1.00)
	Mean (SD)	74 (11.0)	73 (9.8)	75 (13.7)	
	Q25, Median	67, 75	67, 75	67, 78	
	Minimum, maximum	40, 105	40, 103	40, 105	
Sex	Female	18185 (48.0%)	12309 (44.8%)	5876 (56.2%)	Reference
	Male	19722 (52.0%)	15149 (55.2%)	4573 (43.8%)	0.63 (0.61 - 0.66)
Smoking	Non Smoker	16293 (43.0%)	11310 (41.2%)	4983 (47.7%)	Reference
	Ex Smoker	15521 (40.9%)	13288 (48.4%)	2233 (21.4%)	0.38 (0.36 - 0.40)
	Smoker	3611 (9.5%)	2354 (8.6%)	1257 (12.0%)	1.21 (1.13 - 1.29)
	Unknown	2482 (6.5%)	506 (1.8%)	1976 (18.9%)	8.86 (8.10 - 9.63)
Body mass index	Underweight	1997 (5.3%)	1277 (4.7%)	720 (6.9%)	Reference
	Normal	9908 (26.1%)	7403 (27.0%)	2505 (24.0%)	0.60 (0.55 - 0.65)
	Overweight	11695 (30.9%)	9416 (34.3%)	2279 (21.8%)	0.43 (0.39 - 0.47)
	Obese	7466 (19.7%)	6393 (23.3%)	1073 (10.3%)	0.30 (0.27 - 0.33)
	Unknown	6841 (18.0%)	2969 (10.8%)	3872 (37.1%)	2.31 (2.11 - 2.51)

Table 1 (cont.): Baseline characteristics of the warfarin users and AF patients not treated with antithrombotic therapy

		Total (N=37,907)	Warfarin users (N=27,458)	AF patients not treated with antithrombotic therapy (N=10,449)	Crude OR (95% CI)
$CHADS_2$ score	0	3724 (9.8%)	1931 (7.0%)	1793 (17.2%)	Reference
	1	11417 (30.1%)	8140 (29.6%)	3277 (31.4%)	0.43 (0.41 - 0.46)
	2	11230 (29.6%)	8258 (30.1%)	2972 (28.4%)	0.39 (0.36 - 0.41)
	3	6783 (17.9%)	5019 (18.3%)	1764 (16.9%)	0.38 (0.35 - 0.41)
	4	3217 (8.5%)	2779 (10.1%)	438 (4.2%)	0.17 (0.15 - 0.19)
	5 or 6	1536 (4.1%)	1331 (4.8%)	205 (2.0%)	0.17 (0.14 - 0.19)
CHA_2DS_2-VASc score category	Low (0)	1404 (3.7%)	612 (2.2%)	792 (7.6%)	Reference
	Intermediate (1)	3257 (8.6%)	2148 (7.8%)	1109 (10.6%)	0.40 (0.36 - 0.44)
	High (≥2)	33246 (87.7%)	24698 (89.9%)	8548 (81.8%)	0.27 (0.24 - 0.29)
CHA_2DS_2-VASc score	0	1404 (3.7%)	612 (2.2%)	792 (7.6%)	Reference
	1	3257 (8.6%)	2148 (7.8%)	1109 (10.6%)	0.40 (0.36 - 0.44)
	2	5458 (14.4%)	3814 (13.9%)	1644 (15.7%)	0.33 (0.30 - 0.37)
	3	7607 (20.1%)	5399 (19.7%)	2208 (21.1%)	0.32 (0.29 - 0.35)
	4	8439 (22.3%)	6123 (22.3%)	2316 (22.2%)	0.29 (0.26 - 0.32)
	5	6085 (16.1%)	4532 (16.5%)	1553 (14.9%)	0.26 (0.24 - 0.29)
	6	3402 (9.0%)	2839 (10.3%)	563 (5.4%)	0.15 (0.14 - 0.17)
	7	1632 (4.3%)	1427 (5.2%)	205 (2.0%)	0.11 (0.09 - 0.13)
	8	544 (1.4%)	490 (1.8%)	54 (0.5%)	0.09 (0.06 - 0.11)
	9	79 (0.2%)	74 (0.3%)	5 (0.0%)	0.05 (0.01 - 0.09)
	10	0 (0.0%)	0 (0.0%)	0 (0.0%)	

Table 1 (cont.): Baseline characteristics of the warfarin users and AF patients not treated with antithrombotic therapy

		Total (N=37,907)	Warfarin users (N=27,458)	AF patients not treated with antithrombotic therapy (N=10,449)	Crude OR (95% CI)
Medical history	Congestive Heart Failure	8682 (22.9%)	6255 (22.8%)	2427 (23.2%)	1.03 (0.98 - 1.07)[#]
	Diabetes	5211 (13.7%)	4408 (16.1%)	803 (7.7%)	0.44 (0.41 - 0.46)[#]
	Hypertension	29860 (78.8%)	23495 (85.6%)	6365 (60.9%)	0.26 (0.25 - 0.27)[#]
	Stroke or TIA	6253 (16.5%)	5472 (19.9%)	781 (7.5%)	0.32 (0.30 - 0.35)[#]
Medication use in the 6 months before	Aspirin -Low dose (≤100 mg)	10148 (26.8%)	10148 (37.0%)	0 (0.0%)	
	Aspirin - High dose (>100 mg)	705 (1.9%)	705 (2.6%)	0 (0.0%)	
	Clopidogrel	1035 (2.7%)	1035 (3.8%)	0 (0.0%)	
	Anti-arrhythmics	7476 (19.7%)	6495 (23.7%)	981 (9.4%)	0.33 (0.31 - 0.35)[#]
	Rate limiting drugs	21345 (56.3%)	16730 (60.9%)	4615 (44.2%)	0.51 (0.49 - 0.53)[#]

[#]Reference group consists of those without the risk factor of interest

AF: Atrial Fibrillation; $CHADS_2$ and CHA_2DS_2-VASc: clinical prediction rules for estimating the risk of stroke based on Congestive heart failure, Hypertension, Age (≥65 = 1 point, ≥75 = 2 points), Diabetes, Stroke/TIA (2 points), and Vascular disease; CI: Confidence Interval; OR: Odds Ratio; Q25: 25th percentile; SD: Standard Deviation; TIA: Transient Ischaemic Attack.

Table 2: % of time spent within therapeutic range stratified by age, sex and $CHADS_2$ score (for patients with at least three INR measurements)

Category		N patients	Mean % of time		
			Within therapeutic range (SD)	Below therapeutic range (SD)	Above therapeutic range (SD)
All		18113	63.1 (22.0)	20.7 (19.9)	16.2 (17.8)
Age (years)	40-59	1739	61.1 (23.3)	22.1 (22.1)	16.8 (19.4)
	60-69	4154	64.1 (21.5)	19.5 (19.4)	16.4 (18.2)
	70-79	7391	64.3 (21.3)	19.6 (18.7)	16.1 (17.4)
	80+	4829	61.1 (23.0)	22.9 (20.9)	16.0 (17.5)
Sex	Female	7884	62.3 (22.3)	20.8 (19.8)	16.9 (18.2)
	Male	10229	63.7 (21.8)	20.6 (20.0)	15.6 (17.5)
$CHADS_2$ score	0	1364	63.8 (22.3)	21.7 (21.5)	14.5 (16.2)
	1	5627	64.7 (21.4)	19.7 (19.2)	15.6 (17.5)
	2	5393	62.9 (22.0)	21.0 (19.8)	16.1 (17.8)
	3	3164	62.0 (22.1)	20.8 (19.7)	17.2 (18.5)
	4	1748	61.3 (22.6)	21.3 (20.3)	17.4 (18.6)
	5 or 6	817	60.2 (23.3)	22.6 (21.8)	17.2 (18.2)
CHA_2DS_2-VASc score category	Low (0)	444	62.5 (21.8)	24.2 (22.6)	13.3 (14.0)
	Intermediate (1)	1493	63.2 (22.1)	21.3 (21.0)	15.5 (17.4)
	High (≥2)	16176	63.1 (22.0)	20.6 (19.7)	16.3 (17.9)
CHA_2DS_2-VASc score	0	444	62.5 (21.8)	24.2 (22.6)	13.3 (14.0)
	1	1493	63.2 (22.1)	21.3 (21.0)	15.5 (17.4)
	2	2688	64.7 (21.6)	19.7 (19.3)	15.6 (18.0)
	3	3623	64.7 (21.6)	20.0 (19.5)	15.2 (17.2)
	4	3995	62.9 (21.9)	20.6 (19.6)	16.5 (17.8)
	5	2867	62.6 (22.2)	20.6 (19.8)	16.8 (18.1)
	6	1776	61.4 (22.1)	21.2 (19.5)	17.4 (18.2)
	7	883	59.8 (23.2)	22.0 (21.0)	18.2 (19.5)
	8	302	57.0 (23.8)	24.3 (22.7)	18.6 (19.4)
	9	42	56.5 (24.4)	25.3 (23.8)	18.3 (16.6)
	10	-	-	-	-

$CHADS_2$ and CHA_2DS_2-VASc: clinical prediction rules for estimating the risk of stroke based on Congestive heart failure, Hypertension, Age (≥65 = 1 point, ≥75 = 2 points), Diabetes, Stroke/TIA (2 points), and Vascular disease; SD: Standard Deviation.

Table 3: Rate of stroke (including TIA) and % of time spent within therapeutic range in warfarin users compared to AF patients not treated with antithrombotic therapy or to warfarin users with ≤ 30% of time spent within therapeutic range

					Patients with: $CHADS_2$ score ≥ 2			Patients with: CHA_2DS_2-VASc score ≥ 1			Patients with: CHA_2DS_2-VASc score ≥ 2		
Cohort	**%TIR**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**
Non users*		898	Reference	Reference	675	Reference	Reference	891	Reference	Reference	869	Reference	Reference
Warfarin users	≤ 30%	226	4.05 (3.52 - 4.58)	3.80 (3.09 - 4.50)	168	3.13 (2.65 - 3.61)	2.67 (2.07 - 3.27)	225	3.94 (3.42 - 4.46)	3.71 (3.02 - 4.40)	214	3.70 (3.21 - 4.20)	3.39 (2.74 - 4.04)
	31-40%	105	2.34 (1.93 - 2.75)	2.19 (1.71 - 2.68)	80	1.87 (1.49 - 2.25)	1.67 (1.23 - 2.10)	105	2.30 (1.90 - 2.70)	2.19 (1.70 - 2.67)	102	2.20 (1.81 - 2.59)	2.05 (1.59 - 2.52)
	41-50%	149	1.76 (1.50 - 2.03)	1.74 (1.39 - 2.09)	115	1.41 (1.17 - 1.66)	1.30 (0.99 - 1.61)	149	1.73 (1.47 - 1.99)	1.72 (1.38 - 2.07)	143	1.63 (1.38 - 1.88)	1.61 (1.28 - 1.95)
	51-60%	200	1.27 (1.10 - 1.44)	1.31 (1.07 - 1.55)	144	1.02 (0.86 - 1.19)	0.97 (0.75 - 1.18)	200	1.25 (1.08 - 1.41)	1.30 (1.06 - 1.53)	192	1.18 (1.02 - 1.34)	1.21 (0.98 - 1.44)
	61-70%	266	1.00 (0.88 - 1.12)	0.94 (0.78 - 1.10)	198	0.86 (0.73 - 0.98)	0.72 (0.57 - 0.88)	262	0.96 (0.85 - 1.08)	0.91 (0.76 - 1.07)	247	0.90 (0.79 - 1.01)	0.84 (0.69 - 0.99)
	≥ 70%	513	0.90 (0.81 - 0.99)	0.85 (0.72 - 0.98)	364	0.77 (0.68 - 0.86)	0.64 (0.52 - 0.76)	511	0.88 (0.79 - 0.96)	0.83 (0.70 - 0.96)	486	0.83 (0.74 - 0.91)	0.77 (0.65 - 0.89)

Table 3 (cont.): Rate of stroke (including TIA) and % of time spent within therapeutic range in warfarin users compared to AF patients not treated with antithrombotic therapy or to warfarin users with ≤ 30% of time spent within therapeutic range

					Patients with: $CHADS_2$ score ≥ 2			Patients with: CHA_2DS_2-VASc score ≥ 1			Patients with: CHA_2DS_2-VASc score ≥ 2		
Cohort	%TIR	N	Crude RR (95% CI)	Adjusted RR# (95% CI)	N	Crude RR (95% CI)	Adjusted RR# (95% CI)	N	Crude RR (95% CI)	Adjusted RR# (95% CI)	N	Crude RR (95% CI)	Adjusted RR# (95% CI)
Warfarin	≤ 30%	226	Reference	Reference	168	Reference	Reference	225	Reference	Reference	214	Reference	Reference
users	31-40%	105	0.56	0.57	80	0.58	0.60	105	0.57	0.58	102	0.57	0.59
			(0.45 - 0.67)	(0.44 - 0.69)		(0.45 - 0.71)	(0.46 - 0.75)		(0.46 - 0.68)	(0.46 - 0.70)		(0.46 - 0.69)	(0.46 - 0.72)
	41-50%	149	0.42	0.45	115	0.43	0.47	149	0.42	0.46	143	0.42	0.47
			(0.35 - 0.49)	(0.36 - 0.54)		(0.34 - 0.52)	(0.37 - 0.58)		(0.35 - 0.50)	(0.37 - 0.54)		(0.35 - 0.50)	(0.37 - 0.56)
	51-60%	200	0.30	0.33	144	0.31	0.35	200	0.30	0.34	192	0.30	0.35
			(0.25 - 0.35)	(0.28 - 0.39)		(0.25 - 0.37)	(0.28 - 0.43)		(0.25 - 0.35)	(0.28 - 0.40)		(0.25 - 0.35)	(0.28 - 0.41)
	61-70%	266	0.23	0.24	198	0.26	0.26	262	0.23	0.24	247	0.23	0.24
			(0.20 - 0.27)	(0.20 - 0.28)		(0.21 - 0.30)	(0.21 - 0.31)		(0.20 - 0.27)	(0.20 - 0.28)		(0.19 - 0.26)	(0.20 - 0.28)
	≥ 70%	513	0.21	0.21	364	0.23	0.23	511	0.21	0.21	486	0.21	0.22
			(0.18 - 0.24)	(0.18 - 0.25)		(0.19 - 0.27)	(0.19 - 0.27)		(0.18 - 0.24)	(0.18 - 0.25)		(0.18 - 0.24)	(0.18 - 0.25)

* AF patients not treated with antithrombotic therapy.

adjusted for age and sex and for $CHADS_2$ risk score of stroke (without age), smoking history, body mass index, socioeconomic status of practice, alcohol use, duration of AF at baseline and prescribing in the six months before of anti-arrhythmics, heart rate limiting drugs and medications that can interact with warfarin; the comparisons of warfarin users only also included use of other antithrombotics (aspirin, aspirin + dipyridamole, clopidogrel).

AF: Atrial Fibrillation; $CHADS_2$ and CHA_2DS_2-VASc: clinical prediction rules for estimating the risk of stroke based on Congestive heart failure, Hypertension, Age (≥65 = 1 point, ≥75 = 2 points), Diabetes, Stroke/TIA (2 points), and Vascular disease; CI: Confidence Interval; N: Number of cases; RR: Relative Rate; TIA: Transient Ischaemic Attack; %TIR: % time spent within therapeutic range.

Table 4: Rate of stroke (excluding TIA) and percentage of time spent within therapeutic range in warfarin users compared to AF patients not treated with antithrombotic therapy or to warfarin users with ≤ 30% of time spent within therapeutic range

					Patients with: $CHADS_2$ score ≥ 2			Patients with: CHA_2DS_2-VASc score ≥ 1			Patients with: CHA_2DS_2-VASc score ≥ 2		
Cohort	**% TIR**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**
Non users*		742	Reference	Reference	566	Reference	Reference	734	Reference	Reference	719	Reference	Reference
Warfarin users	≤ 30%	143	2.98 (2.50 - 3.46)	3.08 (2.39 - 3.78)	104	2.21 (1.79 - 2.63)	2.09 (1.52 - 2.66)	144	2.95 (2.48 - 3.42)	3.07 (2.38 - 3.76)	133	2.70 (2.26 - 3.15)	2.74 (2.09 - 3.39)
	31-40%	62	1.63 (1.27 - 1.99)	1.65 (1.19 - 2.11)	48	1.31 (0.98 - 1.64)	1.30 (0.87 - 1.72)	62	1.61 (1.25 - 1.97)	1.65 (1.19 - 2.12)	60	1.54 (1.19 - 1.89)	1.56 (1.11 - 2.00)
	41-50%	88	1.23 (0.99 - 1.46)	1.36 (1.02 - 1.70)	64	0.93 (0.72 - 1.14)	0.99 (0.70 - 1.28)	88	1.22 (0.98 - 1.45)	1.35 (1.01 - 1.69)	83	1.13 (0.90 - 1.35)	1.24 (0.92 - 1.57)
	51-60%	127	0.96 (0.80 - 1.12)	1.08 (0.84 - 1.32)	97	0.81 (0.66 - 0.97)	0.86 (0.63 - 1.09)	128	0.96 (0.80 - 1.11)	1.09 (0.85 - 1.33)	120	0.89 (0.74 - 1.04)	1.00 (0.77 - 1.23)
	61-70%	152	0.68 (0.58 - 0.78)	0.67 (0.52 - 0.81)	114	0.58 (0.48 - 0.68)	0.52 (0.38 - 0.66)	148	0.65 (0.55 - 0.75)	0.65 (0.51 - 0.79)	141	0.62 (0.52 - 0.72)	0.60 (0.47 - 0.74)
	≥ 70%	311	0.66 (0.58 - 0.74)	0.68 (0.55 - 0.80)	215	0.55 (0.47 - 0.63)	0.50 (0.38 - 0.62)	310	0.65 (0.57 - 0.72)	0.67 (0.54 - 0.80)	294	0.61 (0.54 - 0.69)	0.62 (0.50 - 0.74)

Table 4 (cont.): Rate of stroke (excluding TIA) and percentage of time spent within therapeutic range in warfarin users compared to AF patients not treated with antithrombotic therapy or to warfarin users with ≤ 30% of time spent within therapeutic range

					Patients with: $CHADS_2$ score ≥ 2			Patients with: CHA_2DS_2-VASc score ≥ 1			Patients with: CHA_2DS_2-VASc score ≥ 2		
Cohort	**% TIR**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**
Warfarin users	≤ 30%	143	Reference	Reference	104	Reference	Reference	144	Reference	Reference	133	Reference	Reference
	31-40%	62	0.52 (0.39 - 0.66)	0.52 (0.38 - 0.66)	48	0.57 (0.40 - 0.73)	0.60 (0.41 - 0.79)	62	0.53 (0.39 - 0.66)	0.53 (0.38 - 0.67)	60	0.55 (0.41 - 0.69)	0.55 (0.40 - 0.71)
	41-50%	88	0.40 (0.31 - 0.48)	0.43 (0.33 - 0.54)	64	0.40 (0.29 - 0.51)	0.46 (0.33 - 0.59)	88	0.40 (0.31 - 0.48)	0.43 (0.33 - 0.54)	83	0.40 (0.30 - 0.49)	0.44 (0.33 - 0.56)
	51-60%	127	0.30 (0.24 - 0.36)	0.34 (0.27 - 0.42)	97	0.35 (0.26 - 0.43)	0.40 (0.30 - 0.51)	128	0.31 (0.24 - 0.37)	0.35 (0.27 - 0.42)	120	0.31 (0.24 - 0.38)	0.36 (0.27 - 0.44)
	61-70%	152	0.21 (0.17 - 0.25)	0.21 (0.16 - 0.25)	114	0.24 (0.19 - 0.30)	0.24 (0.18 - 0.30)	148	0.21 (0.17 - 0.25)	0.20 (0.16 - 0.25)	141	0.21 (0.17 - 0.26)	0.21 (0.16 - 0.26)
	≥ 70%	311	0.20 (0.17 - 0.24)	0.21 (0.17 - 0.25)	215	0.23 (0.18 - 0.28)	0.23 (0.18 - 0.28)	310	0.20 (0.17 - 0.24)	0.21 (0.17 - 0.25)	294	0.21 (0.17 - 0.25)	0.21 (0.17 - 0.26)

* AF patients not treated with antithrombotic therapy.

adjusted for age and sex and for $CHADS_2$ risk score of stroke (without age), smoking history, body mass index, socioeconomic status of practice, alcohol use, duration of AF at baseline and prescribing in the six months before of anti-arrhythmics, heart rate limiting drugs and medications that can interact with warfarin; the comparisons of warfarin users only also included use of other antithrombotics (aspirin, aspirin + dipyridamole, clopidogrel).

AF: Atrial Fibrillation; $CHADS_2$ and CHA_2DS_2-VASc: clinical prediction rules for estimating the risk of stroke based on Congestive heart failure, Hypertension, Age (≥65 = 1 point, ≥75 = 2 points), Diabetes, Stroke/TIA (2 points), and Vascular disease; CI: Confidence Interval; N: Number of cases; RR: Relative Rate; TIA: Transient Ischaemic Attack; %TIR: % time spent within therapeutic range.

Table 5: Rate of mortality and percentage of time spent within therapeutic range in warfarin users compared to AF patients not treated with antithrombotic therapy or to warfarin users with ≤ 30% of time spent within therapeutic range

					Patients with: $CHADS_2$ score ≥ 2			Patients with: CHA_2DS_2-VASc score ≥ 1			Patients with: CHA_2DS_2-VASc score ≥ 2		
Cohort	**% TIR**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**
Non users*		5624	Reference	Reference	3704	Reference	Reference	5554	Reference	Reference	5298	Reference	Reference
Warfarin users	≤ 30%	554	1.34 (1.24 - 1.44)	1.44 (1.29 - 1.59)	450	1.29 (1.18 - 1.40)	1.52 (1.34 - 1.71)	550	1.32 (1.22 - 1.42)	1.43 (1.28 - 1.58)	536	1.31 (1.21 - 1.41)	1.44 (1.29 - 1.60)
	31-40%	312	0.91 (0.82 - 1.00)	0.94 (0.82 - 1.06)	248	0.84 (0.75 - 0.93)	0.94 (0.81 - 1.08)	312	0.91 (0.82 - 0.99)	0.95 (0.83 - 1.07)	303	0.89 (0.81 - 0.98)	0.95 (0.83 - 1.07)
	41-50%	462	0.74 (0.68 - 0.80)	0.78 (0.69 - 0.86)	373	0.68 (0.62 - 0.74)	0.77 (0.68 - 0.87)	459	0.73 (0.67 - 0.79)	0.77 (0.69 - 0.86)	449	0.72 (0.66 - 0.78)	0.77 (0.69 - 0.86)
	51-60%	676	0.56 (0.52 - 0.60)	0.62 (0.56 - 0.68)	536	0.54 (0.50 - 0.58)	0.63 (0.56 - 0.70)	673	0.55 (0.52 - 0.59)	0.61 (0.55 - 0.67)	657	0.55 (0.51 - 0.59)	0.62 (0.56 - 0.68)
	61-70%	803	0.38 (0.36 - 0.40)	0.42 (0.38 - 0.46)	624	0.36 (0.34 - 0.39)	0.44 (0.39 - 0.49)	798	0.38 (0.35 - 0.40)	0.42 (0.38 - 0.46)	780	0.37 (0.35 - 0.40)	0.42 (0.38 - 0.46)
	≥ 70%	1235	0.28 (0.27 - 0.30)	0.33 (0.30 - 0.36)	942	0.28 (0.26 - 0.30)	0.35 (0.31 - 0.38)	1229	0.28 (0.27 - 0.30)	0.33 (0.30 - 0.36)	1188	0.28 (0.26 - 0.29)	0.33 (0.30 - 0.36)

Table 5 (cont.): Rate of mortality and percentage of time spent within therapeutic range in warfarin users compared to AF patients not treated with antithrombotic therapy or to warfarin users with ≤ 30% of time spent within therapeutic range

					Patients with: $CHADS_2$ score ≥ 2			Patients with: CHA_2DS_2-VASc score ≥ 1			Patients with: CHA_2DS_2-VASc score ≥ 2		
Cohort	**% TIR**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**	**N**	**Crude RR (95% CI)**	**Adjusted RR# (95% CI)**
Warfarin users	≤ 30%	554	Reference	Reference	450	Reference	Reference	550	Reference	Reference	536	Reference	Reference
	31-40%	312	0.59 (0.52 - 0.66)	0.59 (0.51 - 0.67)	248	0.57 (0.50 - 0.65)	0.56 (0.47 - 0.64)	312	0.60 (0.53 - 0.67)	0.60 (0.52 - 0.68)	303	0.59 (0.52 - 0.66)	0.59 (0.51 - 0.67)
	41-50%	462	0.47 (0.42 - 0.51)	0.48 (0.42 - 0.54)	373	0.45 (0.40 - 0.50)	0.45 (0.39 - 0.51)	459	0.47 (0.42 - 0.52)	0.48 (0.43 - 0.54)	449	0.46 (0.42 - 0.51)	0.48 (0.42 - 0.54)
	51-60%	676	0.34 (0.31 - 0.37)	0.37 (0.33 - 0.41)	536	0.34 (0.31 - 0.38)	0.36 (0.32 - 0.40)	673	0.34 (0.31 - 0.37)	0.37 (0.33 - 0.41)	657	0.34 (0.31 - 0.37)	0.37 (0.33 - 0.41)
	61-70%	803	0.22 (0.20 - 0.24)	0.24 (0.22 - 0.27)	624	0.22 (0.20 - 0.25)	0.24 (0.21 - 0.27)	798	0.22 (0.20 - 0.24)	0.25 (0.22 - 0.27)	780	0.22 (0.20 - 0.24)	0.24 (0.22 - 0.27)
	≥70%	1235	0.17 (0.15 - 0.18)	0.19 (0.17 - 0.21)	942	0.17 (0.16 - 0.19)	0.19 (0.17 - 0.22)	1229	0.17 (0.15 - 0.18)	0.19 (0.17 - 0.21)	1188	0.17 (0.15 - 0.18)	0.19 (0.17 - 0.21)

* AF patients not treated with antithrombotic therapy.

adjusted for age and sex and for $CHADS_2$ risk score of stroke (without age), smoking history, body mass index, socioeconomic status of practice, alcohol use, duration of AF at baseline and prescribing in the six months before of anti-arrhythmics, heart rate limiting drugs and medications that can interact with warfarin; the comparisons of warfarin users only also included use of other antithrombotics (aspirin, aspirin + dipyridamole, clopidogrel)

AF: Atrial Fibrillation; $CHADS_2$ and CHA_2DS_2-VASc: clinical prediction rules for estimating the risk of stroke based on Congestive heart failure, Hypertension, Age (≥65 = 1 point, ≥75 = 2 points), Diabetes, Stroke/TIA (2 points), and Vascular disease; CI: Confidence Interval; N: Number of cases; RR: Relative Rate; TIA: Transient Ischaemic Attack; %TIR: % time spent within therapeutic range.

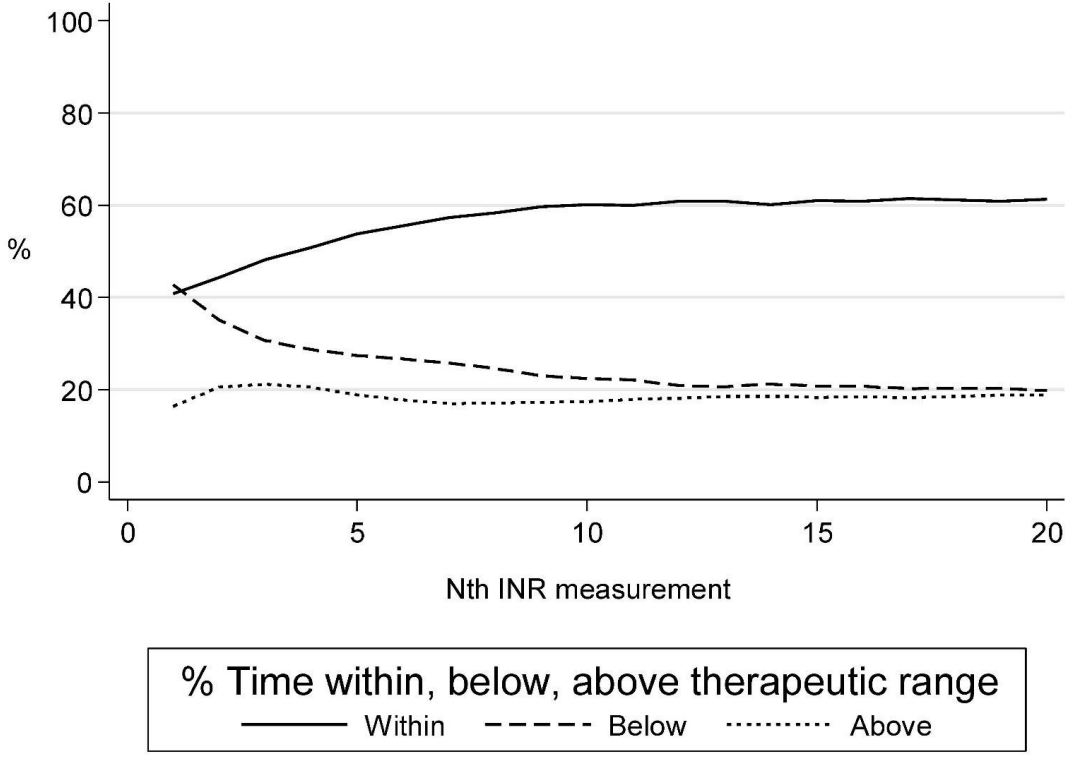


Figure 1: The percentage of INR measurements within (INR 2.0-3.0), below (INR <2.0) or above (INR>3.0) therapeutic range stratified by the number of measurements.

INR: International normalised ratio

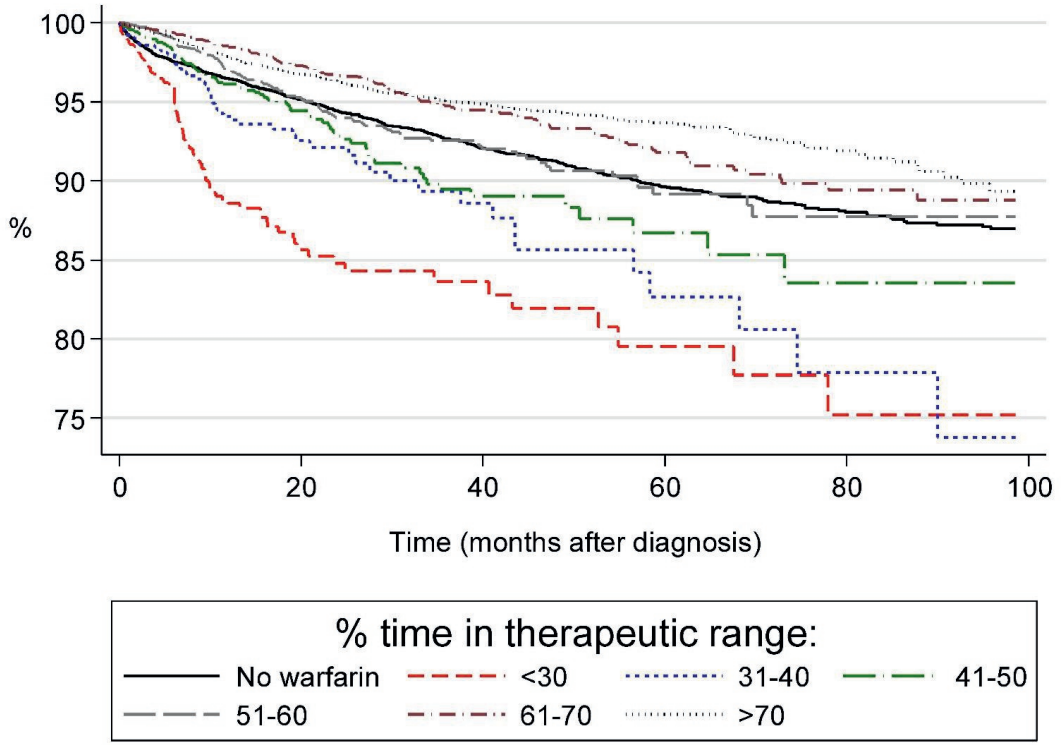


Figure 2: Percentage of patients without a stroke over time stratified by time spent within therapeutic range (INR 2.0-3.0)

INR: International normalised ratio

Discussion

This study found that 63.1% of the time treated with warfarin was spent within therapeutic range. Time in therapeutic range was a strong predictor of the risk of stroke. Decreasing anticoagulation control was generally associated with larger risks of stroke.

Previous research has reported that in high-risk AF patients admitted to hospital with a stroke, and who were candidates for anticoagulation, most were either not taking warfarin or had a subtherapeutic INR at the time of the event [15]. Moreover, several US studies found that the INR values were out of the target range approximately half the time [16, 17]. In the large ACTIVE W clinical trial, it was found that the benefits of oral anticoagulation therapy were most pronounced in patients in centres with a mean percentage of time spent within therapeutic range above 65% [18]. There are multiple and heterogeneous reasons for suboptimal anticoagulation with warfarin. These include, among others, co-prescribing of drugs that can interact with warfarin. The list of drugs with a potential to interact with warfarin is continuously increasing. A recent GPRD study found that the AF patients with the highest risks of stroke were most likely to be prescribed other drugs [10]. More complex dosage instructions may adversely affect patient compliance [19]. The presence of anticoagulation services [15] or patient self-testing of prothrombin time [20] may contribute to improved anticoagulation. Nevertheless, the challenge remains to provide optimal anticoagulation to patients with AF over many years and to improve long-term persistence to anticoagulation therapy [10]. Hopefully, the next generation of oral anticoagulants that do not require anticoagulation monitoring and have less drug-drug and drug-food interactions may help to improve this. For example, in the recently published Randomised Evaluation of Long-Term Anticoagulation Therapy (RE-LY) study, the new oral direct thrombin inhibitor, dabigatran etexilate, provided a lower rate of stroke and systemic embolism than warfarin (1.54% per year, RR of 0.90, $p<0.001$ for non-inferiority with the 110 mg twice-daily dose and 1.11% per year, RR of 0.65, $p<0.001$ for superiority with the 150 mg twice-daily dose versus 1.71% per year with warfarin), and a lower rate of haemorrhagic stroke (0.12% per year, RR of 0.31, $p<0.001$ with the lower dose and 0.10% per year, RR of 0.26, $p<0.001$ with the higher dose versus 0.38% per year with warfarin) in patients with AF [21, 22].

Most warfarin users and most of controls with AF and not treated with antithrombotic therapy in the present study were found to be at high risk of stroke according to the CHA_2DS_2-VASc risk score. But several studies have reported that current risk stratification schemes, including CHA_2DS_2-VASc, have only limited ability to predict the risk of stroke [23-26]. Also, elderly patients with AF are less likely to receive coagulation therapy [10] despite being most at risk of stroke. These findings indicate

that currently that there may be limited evidence to guide the clinicians in deciding who to treat or not with anticoagulation therapy. It has been proposed that pragmatically all AF patients over 75 years should be treated with oral anticoagulation therapy until better prediction tools are available [23]. But not all patients who start oral anticoagulation therapy will respond well (in the present study, 36.9% of patients were outside therapeutic range during follow-up). Thus, there is a need not only to improve our tools for predicting of stroke risk but also of response to oral anticoagulation therapy.

There are several important limitations of this study. We did not have information on the diagnostic criteria used for the AF diagnosis (such as electrocardiograms). However, a previous GPRD study reported a high level of validity in the recording of AF by GPs [27]. A reason for this may be that a diagnosis of AF is not typically made only on clinical symptoms but after further investigation (for example, examination by a cardiologist). We were also unable to classify most stroke cases for the underlying aetiology (ischemia or bleeding) as most strokes were recorded in GPRD without these details. The procedures used to diagnose strokes (such as CT or MRI) are not routinely recorded in GPRD. Another limitation is that we did not have information on all risk factors for stroke, such as frailty. But we were able to estimate the well-validated $CHADS_2$ risk score and the association between this risk score and the rate of stroke was in line with that reported in literature [13, 28]. Another limitation was that aspirin is available in the United Kingdom over the counter for small quantities without the need for a GP prescription. The consequence of this may be that the AF patients not treated with antithrombotic therapy may have been using aspirin which has been shown to reduce the incidence of stroke by 19% [2]. Also, like in any epidemiological study, the comparison groups in this study were not randomised. However, as our main comparisons were within the warfarin group, this limitation should have played a lesser role, as levels of achieved INR control are not influenced by selection bias, unlike in other observational studies. Lastly, a limitation was that we did not have information on INR values for all patients as not all anticoagulation clinics report these electronically to the GPs. It remains unclear whether there is systematic bias in the reporting of INR data. Therefore the mean percentage time spent in therapeutic range for the overall population should be treated with caution.

In conclusion, this study found that good anticoagulation was associated with a reduction in the risk of stroke versus no therapy. Patients with >70% of time spent within therapeutic range had considerable lower risks of stroke compared to patients with <30% in range.

What is known about this topic?

- Atrial fibrillation carries an increased risk of stroke and systemic embolism.
- Warfarin is known to reduce this risk, but the effectiveness depends on the quality of anticoagulation measured via the International Normalised Ratio (INR) as percentage time in range.
- Time in range in routine practice settings is known to vary, affecting the real-life effectiveness of warfarin.

What does this paper add?

- The study was conducted in a real-world setting and quantifies the relationship between quality of INR control (time within therapeutic range) and stroke risk.
- The mean real-world time within therapeutic range in this database was 63%.
- Compared to bad INR control (time in range up to 30%), stroke risk in the group with best INR control (time in range of at least 70%) was reduced by 79%.

References

1. Fuster V, Rydén LE, Cannom DS, Crijns HJ, Curtis AB, Ellenbogen KA, Halperin JL, Le Heuzey J-Y, Kay GN, Lowe JE, Olsson SB, Prystowsky EN, Tamargo JL, Wann S, Smith SC Jr, Jacobs AK, Adams CD, Anderson JL, Antman EM, Hunt SA, Nishimura R, Ornato JP, Page RL, Riegel B, Priori SG, Blanc J-J, Budaj A, Camm AJ, Dean V, Deckers JW, Despres C, Dickstein K, Lekakis J, McGregor K, Metra M, Morais J, Osterspey A, Zamorano JL. ACC/AHA/ESC 2006 guidelines for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines (Writing Committee to Revise the 2001 Guidelines for the Management of Patients With Atrial Fibrillation). Circulation. 2006;114:e257-e354.
2. Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. Ann Intern Med. 2007;146:857-67.
3. Rietbrock S, Plumb JM, Gallagher AM, van Staa TP. How effective are dose-adjusted warfarin and aspirin for the prevention of stroke in patients with chronic atrial fibrillation? An analysis of the UK General Practice Research Database. Thromb Haemost. 2009;101:527-34.
4. Fang MC, Go AS, Hylek EM, *et al*. Age and the risk of warfarin-associated hemorrhage: the anticoagulation and risk factors in atrial fibrillation study. J Am Geriatr Soc. 2006;54:1231-6.
5. Hylek EM, Go AS, Chang Y, *et al*. Effect of intensity of oral anticoagulation on stroke severity and mortality in atrial fibrillation. N Engl J Med. 2003;349:1019-26.
6. van Walraven C, Jennings A, Oake N, *et al*. Effect of study setting on anticoagulation control: a systematic review and metaregression. Chest. 2006;129(5):1155-66.
7. Jones M, McEwan P, Morgan CL, *et al*. Evaluation of the pattern of treatment, level of anticoagulation control, and outcome of treatment with warfarin in patients with non-valvar atrial fibrillation: a record linkage study in a large British population. Heart. 2005;91:472-7.
8. Currie CJ, Jones M, Goodfellow J, *et al*. Evaluation of survival and ischaemic and thromboembolic event rates in patients with non-valvar atrial fibrillation in the general population when treated and untreated with warfarin. Heart. 2006;92:196-200.
9. Morgan CL, McEwan P, Tukiendorf A *et al*. Warfarin treatment in patients with atrial fibrillation: Observing outcomes associated with varying levels of INR control. Thromb Res. 2009;124:37-41.
10. Gallagher AM, Rietbrock S, Plumb J, van Staa TP. Initiation and persistence of warfarin or aspirin in patients with chronic atrial fibrillation in general practice: do the appropriate patients receive stroke prophylaxis? J Thromb Haemost. 2008;6:1500-6.
11. Parkinson J, Davis S, van Staa TP. The General Practice Research (GPRD) Database: Now and the Future. In Pharmacovigilance 2007. Editor R Mann. Second edition. John Wiley & Sons.
12. Rosendaal FR, Cannegieter SC, van der Meer FJ, Briët E. A method to determine the optimal intensity of oral anticoagulant therapy. Thromb Haemost. 1993;69:236-9.
13. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA. 2001;285:2864-70.
14. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach. Chest. 2010;137:263-72.

15. Gladstone DJ, Bui E, Fang J, Laupacis A, Lindsay MP, Tu JV, Silver FL, Kapral MK. Potentially preventable strokes in high-risk patients with atrial fibrillation who are not adequately anticoagulated. Stroke. 2009;40:235-40.
16. Samsa GP, Matchar DB, Goldstein LB, Bonito AJ, Lux LJ, Witter DM, Bian J. Quality of anticoagulation management among patients with atrial fibrillation: results of a review of medical records from 2 communities. Arch Intern Med. 2000;160:967-73.
17. Cohen N, Almoznino-Sarafian D, Alon I, Gorelik O, Koopfer M, Chachashvily S, Shteinshnaider M, Litvinjuk V, Modai D. Adequacy of anticoagulation in patients with atrial fibrillation: effect of various parameters. Clin Cardiol. 2001;24:380-4.
18. Connolly SJ, Pogue J, Eikelboom J, Flaker G, Commerford P, Franzosi MG, Healey JS, Yusuf S; ACTIVE W Investigators. Benefit of oral anticoagulant over antiplatelet therapy in atrial fibrillation depends on the quality of international normalized ratio control achieved by centers and countries as measured by time in therapeutic range. Circulation. 2008;118:2029-37.
19. Hixson-Wallace JA, Dotson JB, Blakey SA. Effect of regimen complexity on patient satisfaction and compliance with warfarin therapy. Clin Appl Thromb Hemost. 2001;7:33-7.
20. Yang DT, Robetorye RS, Rodgers GM. Home prothrombin time monitoring: a literature analysis. Am J Hematol. 2004;77:177-86.
21. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, Pogue J, Reilly PA, Themeles E, Varrone J, Wang S, Alings M, Xavier D, Zhu J, Diaz R, Lewis BS, Darius H, Diener HC, Joyner CD, Wallentin L; RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med. 2009;361:1139-51.
22. Connolly SJ, Ezekowitz MD, Yusuf S, Reilly PA, Wallentin L; Randomized Evaluation of Long-Term Anticoagulation Therapy Investigators. Newly identified events in the RE-LY trial. N Engl J Med. 2010;363:1875-6.
23. Hobbs FD, Roalfe AK, Lip GY, Fletcher K, Fitzmaurice DA, Mant J; on behalf of the Birmingham Atrial Fibrillation in the Aged (BAFTA) investigators and Midland Research Practices Consortium (MidReC) network. Performance of stroke risk scores in older people with atrial fibrillation not taking warfarin: comparative cohort study from BAFTA trial. BMJ. 2011;342:d3653.
24. van Staa TP, Setakis E, Di Tanna GL, Lane DA, Lip GY. A comparison of risk stratification schema for stroke in 79884 atrial fibrillation patients in general practice. J Thromb Haemost. 2010;9(1):39-48.
25. Olesen JB, Lip GY, Hansen ML, Hansen PR, Tolstrup JS, Lindhardsen J, Selmer C, Ahlehoff O, Olsen AM, Gislason GH, Torp-Pedersen C. Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. BMJ. 2011;342:d124.
26. Lip GY, Frison L, Halperin JL, Lane DA. Identifying patients at high risk for stroke despite anticoagulation: a comparison of contemporary stroke risk stratification schemes in an anticoagulated atrial fibrillation cohort. Stroke. 2010;41:2731-8.
27. Ruigómez A, Johansson S, Wallander M-A, García-Rodríguez LA. Incidence of chronic atrial fibrillation in general practice and its treatment pattern. J Clin Epidemiol. 2002;55:358-63.
28. Rietbrock S, Heeley E, Plumb J,van Staa TP. Chronic atrial fibrillation: Incidence, prevalence, and prediction of stroke using the Congestive heart failure, Hypertension, Age >75, Diabetes mellitus, and prior Stroke or transient ischemic attack ($CHADS_2$) risk stratification scheme. Am Heart J. 2008;156:57-64.

Chapter 3.2

Population-based cohort study of warfarin-treated patients with atrial fibrillation: Incidence of cardiovascular and bleeding outcomes

Arlene M. Gallagher, Tjeerd P. van Staa, Tarita Murray-Thomas, Nils Schoof, Andreas Clemens, Diana Ackermann, Dorothee B. Bartels

BMJ Open. 2014;4(1):e003839

Abstract

Objectives: Atrial fibrillation (AF) is the most common cardiac rhythm disorder with significant health burden. The aim of this study was to characterize patients with recently diagnosed atrial fibrillation (AF) and to estimate the rates of comorbidities and outcome events requiring hospitalisation in routine clinical practice.

Design: Pharmacoepidemiological cohort study using observational data.

Methods/Setting: This study included 16,513 patients with a first diagnosis of AF between 1 January 2005 and 28 February 2010 (newly diagnosed patients) using data from the UK Clinical Practice Research Data Link (CPRD) linked to Hospital Episode Statistics (HES) and the Office for National Statistics (ONS) mortality data. Exposure was stratified by Vitamin K Antagonist (VKA) exposure (non-use, current, recent and past exposure) based on prescriptions and/or international normalised ratio measurements, and followed for outcome events of interest based on diagnosis codes in the databases, that is, vascular outcomes, bleeding events and others. The main focus of the study was on outcome events requiring hospitalisation using the HES data.

Results: The incidence of hospitalised vascular outcomes (myocardial infarction (MI), stroke or systemic arterial peripheral embolism) was 3.8 (95% CI 3.5 to 4.0)/100 patient-years. The incidence of stroke was 0.9 (0.8 to 1.1) during current VKA exposure, 2.2 (1.6 to 2.9) for recent, 2.4 (1.9 to 2.9) for past, and 3.4 (3.1 to 3.7) during non-use. MI incidence was 0.7 (0.6 to 0.9) for current VKA exposure, 0.7 (0.4 to 1.2) for recent, 1.1 (0.8 to 1.5) for past and 1.9 (1.7 to 2.1) during non-use. The incidence of bleeding event hospitalisations was 3.8 (3.4 to 4.2) for current VKA exposure, 4.5 (3.7 to 5.5) for recent, 2.7 (2.2 to 3.3) for past and 2.9 (2.6 to 3.2) during non-use; 38% of intracranial bleeds and 6% of gastrointestinal bleeds were fatal.

Conclusion: This population-based study from recent years provides a comprehensive characterisation of newly diagnosed AF patients and incidence estimates of common outcomes with a focus on hospitalised events stratified by VKA exposure. This study will help to place future data on new oral anticoagulants into perspective.

Introduction

Atrial fibrillation (AF) is the most common cardiac rhythm disorder and represents a significant health care burden globally. There are two cornerstones in AF management; thromboprophylaxis and heart rate or heart rhythm control [1]. The leading anticoagulation treatment options are vitamin K antagonists (VKAs). The main challenges of treatment with VKAs are close monitoring of the anticoagulant effect (International Normalised Ratio (INR) measures) to maintain the right anticoagulation level, dietary restrictions to allow for a constant dosing, and several drug–drug interactions. Although VKAs are very effective in stroke prevention when optimally dosed, the management challenges present a significant unmet need for other treatment options [2]. New oral anticoagulants (NOACs) were developed and are available for the prevention of stroke in patients with AF. The reversible direct thrombin inhibitor, dabigatran etexilate, was the first approved NOAC based on the RE-LY trial, and is available in more than 90 countries worldwide [3]. Recently, the factor Xa inhibitors rivaroxaban and apixaban were also approved.

To supplement the data seen in the controlled environment of clinical trials, routine data are important to describe the rates of comorbidities and outcome events in the target population as seen in clinical practice. Such data should originate from the same time frame in which the corresponding clinical trials were conducted, since secular trends including changes in concomitant treatments might influence disease prevalence and incidence. Of particular importance are cardiovascular diseases, which often coexist with AF [4]. In addition, known side effects of standard treatment such as bleeding are of importance.

The objective and uniqueness of this study was to describe the characteristics of patients with a first diagnosis of AF between 1 January 2005 and 28 February 2010 (newly diagnosed patients) and to estimate the incidence of cardiovascular and other outcomes, including bleeding events, in patients with AF, overall and in patients being exposed/non-exposed to VKA in the same time frame of the conduct specifically of the RE-LY trial.

Methods

Data sources

This study used the Clinical Practice Research Datalink (CPRD); their primary care database (GOLD), the Hospital Episode Statistics (HES) and the Office for National Statistics (ONS) mortality data from the UK.

The CPRD GOLD (formally known as General Practice Research Database (GPRD)) database comprises computerised medical records of general practitioners (GPs) from the UK. Patients are affiliated with a practice, which centralises the medical information from the GPs, specialist referrals, and hospitalisations. The data recorded includes demographic information, prescriptions, clinical events, preventive care, specialist referrals, hospital admissions and their major outcomes [5]. CPRD currently includes about 8% of the UK population. A recent systematic review of all validation studies found that medical data in CPRD GOLD were generally of high validity; this included an assessment of AF and myocardial infarction (MI) [6]. Practices contributing to the database are each assigned a date at which they are considered up-to-standard (UTS). The CPRD recommend that analyses are performed on data following the practice UTS date.

The national Hospital Episode Statistics (HES) data contain details of all admissions to National Health Service (NHS) hospitals in England. For each hospitalised patient, the hospital charts are reviewed, dates of admission and discharge and main diagnoses are extracted, coded by coding staff and collated nationally into HES. Further details on HES data can be found in the supporting information. Data are available from April 1997. The ONS mortality data contain the date and cause of death, from the official death certificate, for the population of England and Wales from January 2001. HES and ONS data are linked to GOLD data via the NHS number; a unique identifier. The linkage uses a combination of the patient's NHS number, gender, and partial date of birth. Not all patients in GOLD are eligible to be linked to these data sources, for example if they reside in Scotland, or they lack a valid NHS identifier. Over 40% of patients in GOLD have now been linked individually and anonymously to the HES and to the ONS mortality data.

Study population

Patients were included in the study if their first AF record occurring during the study period (1 January 2005–28 February 2010) was identified in GOLD, where they had at least one year of UTS follow-up prior to the first AF diagnosis (i.e. incident patients). The Read codes used for defining AF can be found in Table S4 (supporting information). No free text or further information was used to identify/validate the diagnosis of AF. The index date was the date of this first AF record. Patients were

excluded if they were aged less than 18 at the index date, or if they had a history of heart valve problems/replacement prior. This population was further restricted to those who were eligible for linkage with HES and ONS mortality data (i.e. have a valid NHS number and are registered at a practice that was participating in the linkage programme), creating the study population of interest (Figure 1). In Table S5 (supporting information), a comparison between the incident AF population in CPRD and the incident AF population eligible for linkage can be found. Patients were followed from the index date until the earliest of the respective outcome of interest, transfer out of the practice, last data collection, or death, whichever date came first.

The total period of follow-up was analysed for prescriptions for VKA and/or laboratory measurements for INR as a proxy for VKA treatment. The duration of each anticoagulation prescription was taken as 28 days. Patients who received a repeat prescription within 30 days after the expected end of a prescription were assumed to be continuously treated. INR measurements were treated as an indicator for VKA exposure and therefore treated in the same way as prescriptions (i.e. assigning 28 days of treatment). The follow-up period for each patient was divided into periods of non-use, current, recent and past exposure:

i. Non-use: the time between diagnosis of AF and the first prescription of an anticoagulant/INR measurement (if they started treatment after diagnosis), or the whole follow-up for patients that never receive treatment;
ii. Current exposure: the period when there is a valid prescription/INR measurement; from the date of the prescription/INR measurement until day 28 after the prescription/INR measurement;
iii. Recent exposure: the period between end of treatment and three months later; from the day after treatment ends (28 days after the last prescription/INR measurement) to 90 days later;
iv. Past exposure: subsequent follow-up time after recent treatment (>90 days after the end of treatment.

In the outcome analysis patients could be assigned different exposure status. Throughout the follow-up time, a single patient could provide person-time to one or more periods of non-use, current, recent or past exposure over the course of the study period, i.e. the categories of exposure are not mutually exclusive on patient level.

Outcomes of interest

For each outcome, we identified the first-ever record during follow-up after diagnosis of AF as defined above. Assessment of the incidence rate of AF in the general practice setting was not in the scope of this study. Patients who experienced the first-ever outcome prior to the start of follow-up were excluded in the analysis of that

outcome (although such patients could be included into the analyses of other outcomes). Four main sets of outcomes were examined: vascular outcomes, bleeding events, hospitalisation due to any cause and due to AF, and all-cause mortality. The vascular outcomes recorded as hospital admission in HES: stroke (categorised into haemorrhagic, ischaemic or unspecified stroke), systemic peripheral arterial embolism, and MI. Fatal events were those where the date of death from the death certificate was within 10 days of the outcome. Bleedings (from HES) were defined based on a comprehensive list of bleeding codes (96 International classification of Diseases (ICD)-10 codes – listed in Table S3, supporting information) and categorised into intracranial and extracranial. Gastrointestinal (GI) bleeds (based on 23 ICD-10 codes) were analysed as a subgroup of extracranial bleeds. Hospitalisation in general and due to AF was assessed using the primary ICD-10 code I48 from the hospital diagnosis. Finally, all-cause mortality was based on the data from death certificates (from ONS).

Covariables

3.2

Covariables were assessed from before or on index date. A medical history (at any time prior to the index date) of acute MI, stroke, systemic peripheral arterial embolism, vascular event, coronary artery disease (CAD), congestive heart failure (CHF), hypertension, diabetes mellitus and liver and renal impairment was recorded. The baseline $CHADS_2$ and CHA_2DS_2 -VASc were also calculated [7, 8]. The $CHADS_2$ score is based on a point system where two points are assigned to a history of stroke or transient ischaemic attack and one point each is assigned for CHF, history of hypertension, age of 75 years or older, and diabetes mellitus [7]. The CHA_2DS_2 -VASc score extends this scoring system by age 65-74, female sex, and history of vascular disease [8]. A recent publication of chronic atrial fibrillation patients showed that the $CHADS_2$ score is predictive for the risk of stroke in the CPRD database [9]. Smoking status was calculated from CPRD records of smoking status and from searching for smoking records in the patient's history using a Read code list [10]. Only records from before or on index date were included. Records were categorised into non-smoker, ex-smoker, and smoker.

Prescriptions issued during the three months before the index date of anti-arrhythmics, statins, antiplatelets, anticoagulants, diabetes medication and antihypertensives were also captured. Antihypertensives included angiotensin-converting enzyme inhibitors, angiotensin-II receptor antagonists, beta-adrenoceptor blocking drugs, calcium channel blockers, centrally acting antihypertensive drugs, diuretics with potassium, thiazides and related diuretics.

Deprivation was calculated mapped the postcode of the patient to the Index of Multiple Deprivation (IMD). The IMD combines a number of indicators into a single deprivation score for each small area. The indicators between regions in the UK, but

generally include income, employment, health, and the living environment. The quintiles relate to the country as a whole.

Sensitivity analyses

Several sensitivity analyses were performed. First, we repeated the analysis on (1) cardiovascular outcomes based on the Read codes recorded by the general practitioners, and (2) all bleeding outcomes (intracranial, extracranial, GI) based on 59 previously published ICD-10 codes [11, 12]. Second, we changed the time window for fatal events from 10 to 30 days, i.e. the date of death from the death certificate was within 30 days of the outcome.

Statistical analyses

Incidence rates per 100 person-years (PY) were estimated and corresponding 95% CI were calculated assuming a Poisson distribution of events. The denominator for rates calculation was derived as the sum of the PY of follow-up of patients during the study period. Person-time for treatment periods consisted of the sum of follow-up accrued on the index date and during probable use of anticoagulation treatment. Use of anticoagulation treatment was defined as the period during prescribing of anticoagulation therapy and/or laboratory measurements for INR. Follow-up was censored at the earliest of the outcome of interest, transfer out of the practice, last data collection, or death. Person-time for untreated periods consisted of the sum of follow-up accrued on the index date and during the earliest of the start of a treatment of interest (minus 1 day), the outcome of interest, transfer out of the practice, last data collection, or death.

All analyses have been performed with Stata version 11.2.

Results

A total of 48,121 patients were identified with a first AF record during UTS follow-up and within the study period. Figure 1 describes how patients were subsequently excluded from the study cohort. The overall cohort of incident AF patients included 42,008 patients, of whom 25,495 were excluded for not being eligible for linkage to the HES or ONS data sources. A total of 16,513 patients were eligible for linkage to HES and ONS mortality data and were the main subject of this study. There were no major differences in patient characteristics between the overall cohort of incident AF patients (N=42,008) and the final study population (N=16,513) (Table S5, supporting information), other than a shorter follow-up in the study population: The median (range) duration of follow-up in the overall cohort was 2.2 (0-6) vs 1.7 (0-5) years in the final linked AF cohort.

Table 1 shows the baseline characteristics for the study population. The mean age (SD) was 77 (11) for females and 72 (12) for males; 48% of the patients were women. Defining the risk of stroke based on $CHADS_2$ score resulted in 21% being at low risk ($CHADS_2$ score of 0), 61% at moderate risk ($CHADS_2$ score of 1 or 2) and 18% at high risk ($CHADS_2$ score of greater than 2).

Overall, 1,151 vascular events (136 fatal events and 1015 non-fatal events) were recorded (Table 2). There were 716 strokes (94 fatal), 68 systemic peripheral arterial embolisms (4 fatal) and 403 MIs (43 fatal). The incidence of vascular outcomes overall was 3.8 (95% CI 3.5 to 4.0) per 100 PY, and was 1.8 (95% CI 1.6 to 2.1) during current exposure, 3.1 (2.4 to 3.9) for recent, 3.5 (2.9 to 4.2) for past exposure, and 5.5 (5.1 to 5.9) during non-use. The incidence of stroke overall was 2.3 (2.1 to 2.5), and 0.9 (0.8 to 1.1) during current exposure, 2.2 (1.6 to 2.9) for recent, 2.4 (1.9 to 2.9) for past exposure, and 3.4 (3.1 to 3.7) during non-use. The incidence of MI overall was 1.3 (1.2 to 1.4) and was higher during past- and non-use compared to current and recent exposure (1.1, 1.9 vs 0.7, 0.7). Systemic peripheral arterial embolisms were few (0.2/100 PY). Owing to the low number of fatal events, there was no clear trend with VKA exposure and the incidences of any fatal vascular event or incidences of a fatal stroke, MI or systemic peripheral arterial embolism.

Bleeding events occurred with an incidence of 3.3 (3.1 to 3.5) (Table 3). The incidence rate during current exposure was 3.8 (3.4 to 4.2); and 4.5 (3.7 to 5.5) for recent, whereas non-use and past exposure to VKA resulted in similar incidences (2.9 (2.6 to 3.2) and 2.7 (2.2 to 3.3)). Twelve per cent (n=120) were intracranial bleeds and overall nearly 38% of them were fatal; 50% were fatal during current exposure. More than one third of the extra-cranial bleeds were GI bleeds (n=334), 6% of them were fatal; during current exposure 4% were fatal. GI bleed events occurred similarly frequently during current/recent and past/non-use of VKAs.

The incidences of hospitalisation for any cause, hospitalisation for AF, as well as all-cause mortality are displayed in Table 4. Overall 71% of the patients were hospitalised with an incidence of 88.0 (86.4 to 89.6)/100 person-years for any hospitalisation and 16.0 (15.5 to 16.5) for hospitalisation due to AF.

All-cause mortality was 8.9 (8.6 to 9.3)/100 person-years overall and was highest during recent VKA exposure and lowest during current exposure (15.3 and 3.4, respectively).

The sensitivity analysis for the cardiovascular events based on diagnosis recorded by the GP showed slightly higher incidences than those based on hospital data for MI and stroke, and slightly lower incidences for systemic peripheral arterial embolism (Table S1, supporting information). Bleeding events defined by previously published ICD codes lead to lower bleeding incidences (overall 2.5 (2.3 to 2.7)) and smaller differences between current (3.0 (2.6 to 3.3)) and recent exposure (3.4 (2.7 to 4.3)) (Table S2, supporting information). Changing the time window for fatal events from 10 to 30 days between the outcome of interest and the death certificate date, did not change the results substantially (data not shown).

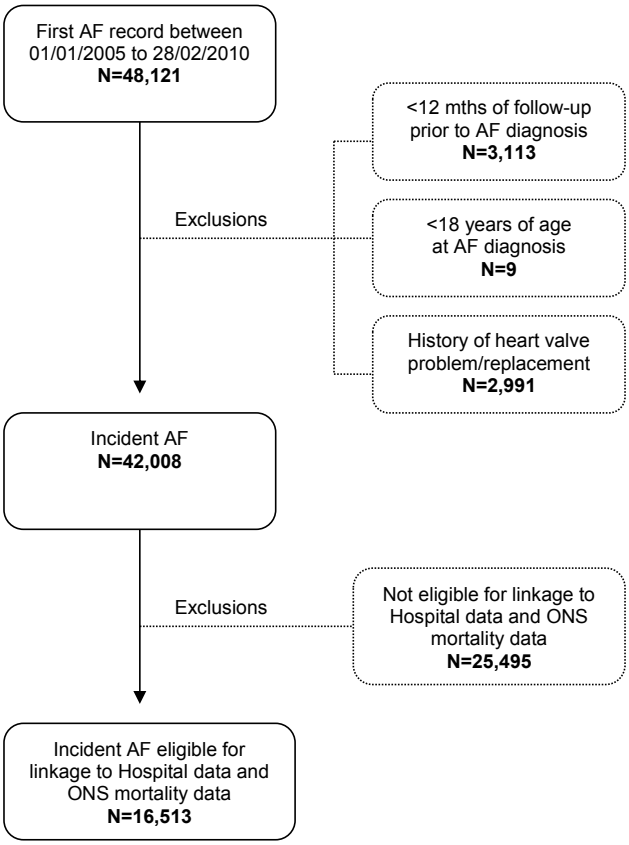


Figure 1: Consort diagram showing the composition of the study population - Incident atrial fibrillation Cohort.

AF: atrial fibrillation; ONS: Office for National Statistics

Table 1: Description of the study population, UK 2005-2010

Characteristics	Female N=7,921	Male N=8,592	Total N=16,513
Length of follow-up (years):			
mean (SD)	1.9 (1.4)	2.0 (1.4)	1.9 (1.4)
median (range)	1.7 (0 - 5)	1.8 (0 - 5)	1.7 (0 - 5)
Age:			
mean (SD)	77 (11.2)	72 (12.2)	74 (12.0)
median (range)	79 (25 - 107)	73 (21 - 101)	76 (21 - 107)
Gender: female	7,921 (100.0%)	-	7,921 (48.0%)
Body Mass Index (kg/m^2):			
mean (SD)	27 (6)	28 (5)	28 (6)
median (range)	26 (12 - 69)	27 (12 - 70)	27 (12 - 70)
Smoker	776 (9.8%)	1,227 (14.3%)	2,003 (12.1%)
Drinker	4,858 (61.3%)	6,743 (78.5%)	11,601 (70.3%)
Deprivation quintile, based on patient's home			
0 (least deprived)	1,251 (15.8%)	1,502 (17.5%)	2,753 (16.7%)
1	1,388 (17.5%)	1,508 (17.6%)	2,896 (17.5%)
2	1,259 (15.9%)	1,376 (16.0%)	2,635 (16.0%)
3	1,219 (15.4%)	1,232 (14.3%)	2,451 (14.8%)
4 (most deprived)	1,144 (14.4%)	1,082 (12.6%)	2,226 (13.5%)
Deprivation level unknown	1,660 (21.0%)	1,892 (22.0%)	3,552 (21.5%)
$CHADS_2$ risk score:			
mean (SD)	1.64 (1.15)	1.37 (1.20)	1.50 (1.18)
median (IQR)	2 (1-2)	1 (0-2)	1 (1-2)
0	1,215 (15.3%)	2,278 (26.5%)	3,493 (21.2%)
1	2,633 (33.2%)	2,875 (33.5%)	5,508 (33.4%)
2	2,574 (32.5%)	2,031 (23.6%)	4,605 (27.9%)
3	897 (11.3%)	881 (10.3%)	1,778 (10.8%)
4	463 (5.8%)	383 (4.5%)	846 (5.1%)
5	131 (1.7%)	129 (1.5%)	260 (1.6%)
6	8 (0.1%)	15 (0.2%)	23 (0.1%)
CHA_2DS_2-VASc risk score:			
mean (SD)	3.74 (1.57)	2.40 (1.7)	3.04 (1.77)
median (IQR)	4 (3-4)	2 (1-3)	3 (2-4)
0	0 (0.0%)	1,258 (14.6%)	1,258 (7.6%)
1	569 (7.2%)	1,405 (16.4%)	1,974 (12.0%)
2	900 (11.4%)	2,199 (25.6%)	3,099 (18.8%)
3	2,219 (28.0%)	1,830 (21.3%)	4,049 (24.5%)
4	2,369 (29.9%)	865 (10.1%)	3,234 (19.6%)
5	788 (9.9%)	502 (5.8%)	1,290 (7.8%)
6	481 (6.1%)	380 (4.4%)	861 (5.2%)
7	451 (5.7%)	136 (1.6%)	587 (3.6%)
8	136 (1.7%)	17 (0.2%)	153 (0.9%)
9	8 (0.1%)	0 (0.0%)	8 (<0.1%)

Table 1 (cont.): Description of the study population, UK 2005-2010

Characteristics	Female N=7,921	Male N=8,592	Total N=16,513
History of*:			
Acute myocardial infarction	515 (6.5%)	1,106 (12.9%)	1,621 (9.8%)
Stroke	884 (11.2%)	960 (11.2%)	1,844 (11.2%)
Systemic peripheral arterial embolism	28 (0.4%)	21 (0.2%)	49 (0.3%)
Vascular event	1,349 (17.0%)	1,924 (22.4%)	3,273 (19.8%)
Coronary artery disease	1,021 (12.9%)	1,755 (20.4%)	2,776 (16.8%)
Congestive heart failure	659 (8.3%)	698 (8.1%)	1,357 (8.2%)
Hypertension	4,280 (54.0%)	3,833 (44.6%)	8,113 (49.1%)
Diabetes	1,098 (13.9%)	1,428 (16.6%)	2,526 (15.3%)
Liver Impairment	133 (1.7%)	204 (2.4%)	337 (2.0%)
Renal Impairment	1,342 (16.9%)	1,171 (13.6%)	2,513 (15.2%)
Treatment with†:			
Anti-arrhythmics	716 (9.0%)	706 (8.2%)	1,422 (8.6%)
Statins	2,435 (30.7%)	3,287 (38.3%)	5,722 (34.7%)
Antiplatelets	3,856 (48.7%)	4,292 (50.0%)	8,148 (49.3%)
Aspirin	3,570 (45.1%)	4,029 (46.9%)	7,599 (46.0%)
Low dose aspirin (<100mg)	3,460 (43.7%)	3,871 (45.1%)	7,331 (44.4%)
High dose aspirin (100mg+)	155 (2.0%)	191 (2.2%)	346 (2.1%)
Clopidogrel	387 (4.9%)	429 (5.0%)	816 (4.9%)
Other anti-platelets	130 (1.6%)	179 (2.1%)	309 (1.9%)
Anticoagulants	762 (9.6%)	1,054 (12.3%)	1,816 (11.0%)
Anti-diabetics	636 (8.0%)	862 (10.0%)	1,498 (9.1%)
Oral anti-diabetics	548 (6.9%)	753 (8.8%)	1,301 (7.9%)
Insulin	158 (2.0%)	207 (2.4%)	365 (2.2%)
Other injectable anti-diabetics	0 (0.0%)	2 (<0.1%)	2 (<0.1%)
Antihypertensives‡	5,361 (67.7%)	5,386 (62.7%)	10,747 (65.1%)

* "Ever" diagnoses of comorbidities were measured prior to the index date in CPRD

† Treatments are measured in the 90 days prior to index date

‡ Antihypertensives include Angiotensin-converting enzyme inhibitors, Angiotensin-II receptor antagonists, Beta-adrenoceptor blocking drugs, Calcium channel blockers, Centrally-acting antihypertensive drugs, Diuretics with potassium, Thiazides and related diuretics

BMI: body mass index; $CHADS_2$ and CHA_2DS_2-VASc: clinical prediction rules for estimating the risk of stroke based on Congestive heart failure, Hypertension, Age (≥65 = 1 point, ≥75 = 2 points), Diabetes, Stroke/TIA (2 points), and Vascular disease; IQR: Interquartile range; SD: standard deviation.

Table 2: Incidence of vascular outcomes overall and in relation to VKA use for the cohort of AF patients eligible for linkage, n=16,513

	Overall			Current Exposure			Recent Exposure			Past Exposure			No use		
Event	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**
Vascular	1151	30622	3.8 (3.5 to 4.0)	205	11137	1.8 (1.6 to 2.1)	69	2243	3.1 (2.4 to 3.9)	128	3644	3.5 (2.9 to 4.2)	749	13599	5.5 (5.1 to 5.9)
Fatal	136	30622	0.4 (0.4 to 0.5)	28	11137	0.3 (0.2 to 0.4)	16	2243	0.7 (0.4 to 1.2)	23	3644	0.6 (0.4 to 0.9)	69	13599	0.5 (0.4 to 0.6)
Non-fatal	1015	30622	3.3 (3.1 to 3.5)	177	11137	1.6 (1.4 to 1.8)	53	2243	2.4 (1.8 to 3.1)	105	3644	2.9 (2.4 to 3.5)	680	13599	5.0 (4.6 to 5.4)
MI	403	31421	1.3 (1.2 to 1.4)	83	11496	0.7 (0.6 to 0.9)	17	2317	0.7 (0.4 to 1.2)	41	3749	1.1 (0.8 to 1.5)	262	13859	1.9 (1.7 to 2.1)
Fatal	43	31421	0.1 (0.1 to 0.2)	6	11496	†	5	2317	0.2 (0.1 to 0.5)	8	3749	0.2 (0.1 to 0.4)	24	13859	0.2 (0.1 to 0.3)
Non-fatal	360	31421	1.1 (1.0 to 1.3)	77	11496	0.7 (0.5 to 0.8)	12	2317	0.5 (0.3 to 0.9)	33	3749	0.9 (0.6 to 1.2)	238	13859	1.7 (1.5 to 1.9)
Stroke	716	31189	2.3 (2.1 to 2.5)	107	11340	0.9 (0.8 to 1.1)	50	2291	2.2 (1.6 to 2.9)	88	3725	2.4 (1.9 to 2.9)	471	13833	3.4 (3.1 to 3.7)
Haemorrhagic	66	31888	0.2 (0.2 to 0.3)	23	11658	0.2 (0.1 to 0.3)	10	2356	0.4 (0.2 to 0.8)	4	3821	†	29	14054	0.2 (0.1 to 0.3)
Ischemic	469	31441	1.5 (1.4 to 1.6)	56	11440	0.5 (0.4 to 0.6)	29	2309	1.3 (0.8 to 1.8)	59	3760	1.6 (1.2 to 2.0)	325	13932	2.3 (2.1 to 2.6)
Unspecified	181	31738	0.6 (0.5 to 0.7)	28	11569	0.2 (0.2 to 0.3)	11	2344	0.5 (0.2 to 0.8)	25	3804	0.7 (0.4 to 1.0)	117	14020	0.8 (0.7 to 1.0)

Table 2 (cont.): Incidence of vascular outcomes overall and in relation to VKA use for the cohort of AF patients eligible for linkage, n=16,513

	Overall			Current Exposure			Recent Exposure			Past Exposure			No use		
Event	N	PY	Incidence* (95% CI)	N	PY	Incidence* (95% CI)	N	PY	Incidence* (95% CI)	N	PY	Incidence* (95% CI)	N	PY	Incidence* (95% CI)
Fatal Stroke	94	31189	0.3 (0.2 to 0.4)	21	11340	0.2 (0.1 to 0.3)	12	2291	0.5 (0.3 to 0.9)	16	3725	0.4 (0.2 to 0.7)	45	13833	0.3 (0.2 to 0.4)
Haemorrhagic	27	31888	0.1 (0.1 to 0.1)	12	11658	0.1 (0.1 to 0.2)	5	2356	0.2 (0.1 to 0.5)	3	3821	†	7	14054	†
Ischemic	44	31441	0.1 (0.1 to 0.2)	7	11440	†	4	2309	†	7	3760	0.2 (0.1 to 0.4)	26	13932	0.2 (0.1 to 0.3)
Unspecified	23	31738	0.1 (<0.1 to 0.1)	2	11569	†	3	2344	†	6	3804	0.2 (0.1 to 0.3)	12	14020	0.1 (<0.1 to 0.1)
Non-fatal Stroke	622	31189	2.0 (1.8 to 2.2)	86	11340	0.8 (0.6 to 0.9)	38	2291	1.7 (1.2 to 2.3)	72	3725	1.9 (1.5 to 2.4)	426	13833	3.1 (2.8 to 3.4)
Haemorrhagic	39	31888	0.1 (0.1 to 0.2)	11	11658	0.1 (<0.1 to 0.2)	5	2356	0.2 (0.1 to 0.5)	1	3821	†	22	14054	0.2 (0.1 to 0.2)
Ischemic	425	31441	1.4 (1.2 to 1.5)	49	11440	0.4 (0.3 to 0.6)	25	2309	1.0 (0.7 to 1.6)	52	3760	1.4 (1.0 to 1.8)	299	13932	2.1 (1.9 to 2.4)
Unspecified	158	31738	0.5 (0.4 to 0.6)	26	11569	0.2 (0.1 to 0.3)	8	2344	0.3 (0.1 to 0.7)	19	3804	0.5 (0.3 to 0.8)	105	14020	0.7 (0.6 to 0.9)
SPAE	68	31855	0.2 (0.2 to 0.3)	21	11617	0.2 (0.1 to 0.3)	6	2351	0.3 (0.1 to 0.6)	6	3823	0.2 (0.1 to 0.3)	35	14063	0.2 (0.2 to 0.3)
Fatal	4	31855	†	3	11617	†	1	2351	†	0	3823	†	0	14063	†
Non-fatal	64	31855	0.2 (0.2 to 0.3)	18	11617	0.2 (0.1 to 0.2)	5	2351	0.2 (0.1 to 0.5)	6	3823	0.2 (0.1 to 0.3)	35	14063	0.2 (0.2 to 0.3)

* Number of cases per 100 person years

† Not enough events to calculate

AF: atrial fibrillation; CI: confidence interval; MI: myocardial infarction; N: number of outcome events; PY: person-years; SPAE: systemic pulmonary arterial embolism; VKA: vitamin K antagonist.

Table 3: Incidence of bleeding outcomes overall and in relation to VKA use for the cohort of AF patients eligible for linkage, n=16,513

	Overall			Current Exposure			Recent Exposure			Past Exposure			No use		
Event	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**
All Bleed	1023	30693	3.3 (3.1 to 3.5)	426	11236	3.8 (3.4 to 4.2)	101	2238	4.5 (3.7 to 5.5)	97	3573	2.7 (2.2 to 3.3)	399	13645	2.9 (2.6 to 3.2)
Intracranial	120	31854	0.4 (0.3 to 0.5)	46	11654	0.4 (0.3 to 0.5)	15	2353	0.6 (0.4 to 1.1)	8	3811	0.2 (0.1 to 0.4)	51	14036	0.4 (0.3 to 0.5)
Extra-cranial	903	30778	2.9 (2.7 to 3.1)	380	11246	3.4 (3.0 to 3.7)	86	2245	3.8 (3.1 to 4.7)	89	3593	2.5 (2.0 to 3.0)	348	13695	2.5 (2.3 to 2.8)
Gastrointestinal	334	31545	1.1 (0.9 to 1.2)	104	11560	0.9 (0.7 to 1.1)	35	2330	1.5 (1.0 to 2.1)	38	3733	1.0 (0.7 to 1.4)	157	13921	1.1 (1.0 to 1.3)
Fatal Bleed	74	30693	0.2 (0.2 to 0.3)	33	11236	0.3 (0.2 to 0.4)	10	2238	0.4 (0.2 to 0.8)	7	3573	0.2 (0.1 to 0.4)	24	13645	0.2 (0.1 to 0.3)
Intracranial	45	31854	0.1 (0.1 to 0.2)	23	11654	0.2 (0.1 to 0.3)	6	2353	0.3 (0.1 to 0.6)	5	3811	0.1 (<0.1 to 0.3)	11	14036	0.1 (<0.1 to 0.1)
Extra-cranial	29	30778	0.1 (0.1 to 0.1)	10	11246	0.1 (<0.1 to 0.2)	4	2245	0.2 (<0.1 to 0.5)	2	3593	0.1 (<0.1 to 0.2)	13	13695	0.1 (0.1 to 0.2)
Gastrointestinal	20	31545	0.1 (<0.1 to 0.1)	4	11560	0.0 (<0.1 to 0.1)	3	2330	0.1 (<0.1 to 0.4)	2	3733	0.1 (<0.1 to 0.2)	11	13921	0.1 (<0.1 to 0.1)
Non-fatal Bleed	949	30693	3.1 (2.9 to 3.3)	393	11236	3.5 (3.2 to 3.9)	91	2238	4.1 (3.3 to 5.0)	90	3573	2.5 (2.0 to 3.1)	375	13645	2.7 (2.5 to 3.0)
Intracranial	75	31854	0.2 (0.2 to 0.3)	23	11654	0.2 (0.1 to 0.3)	9	2353	0.4 (0.2 to 0.7)	3	3811	0.1 (<0.1 to 0.2)	40	14036	0.3 (0.2 to 0.4)
Extra-cranial	874	30778	2.8 (2.7 to 3.0)	370	11246	3.3 (3.0 to 3.6)	82	2245	3.7 (2.9 to 4.5)	87	3593	2.4 (1.9 to 3.0)	335	13695	2.4 (2.2 to 2.7)
Gastrointestinal	314	31545	1.0 (0.9 to 1.1)	100	11560	0.9 (0.7 to 1.1)	32	2330	1.4 (0.9 to 1.9)	36	3733	1.0 (0.7 to 1.3)	146	13921	1.0 (0.9 to 1.2)

* Number of cases per 100 person years

AF: atrial fibrillation; CI: confidence interval; N: number of outcome events; PY: person-years; VKA: vitamin K antagonists.

Table 4: Incidence of hospitalisations and all-cause mortality stratified by VKA use

	Overall			Current Exposure			Recent Exposure			Past Exposure			No use		
Event	N	PY	Incidence* (95% CI)	N	PY	Incidence* (95% CI)	N	PY	Incidence* (95% CI)	N	PY	Incidence* (95% CI)	N	PY	Incidence* (95% CI)
Hospitalisation															
Any	11701	13291	88.0 (86.4 to 89.6)	3110	4601	67.6 (65.2 to 70.0)	599	788	76.0 (70.0 to 82.4)	836	1074	77.8 (72.7 to 83.3)	7156	6827	104.8 (102.4 to 107.3)
AF	3962	24761	16.0 (15.5 to 16.5)	1358	8487	16.0 (15.2 to 16.9)	233	1610	14.5 (12.7 to 16.5)	215	2649	8.1 (7.1 to 9.3)	2156	12015	17.9 (17.2 to 18.7)
Mortality	2858	31939	8.9 (8.6 to 9.3)	396	11663	3.4 (3.1 to 3.7)	361	2359	15.3 (13.8 to 17.0)	440	3830	11.5 (10.4 to 12.6)	1661	14086	11.8 (11.2 to 12.4)

* Number of cases per 100 person years

AF: atrial fibrillation; CI: confidence interval; N: number of outcome events; PY: person-years; VKA: vitamin K antagonists.

Discussion

This study presents a detailed picture of the prevalence and incidence of common comorbidities and outcome events in a large cohort of newly diagnosed patients with AF shortly prior to launch of new NOACs. Moreover, VKA use was stratified into current, recent, past exposure and periods of non-use, showing substantial differences between those exposure groups. Another hallmark of this study is that only newly diagnosed patients were included instead of a mixture of incident and prevalent patients reducing the potential for bias and reflecting the most current population diagnosed with AF.

Vascular event rates were lowest during current VKA exposure and showed increasing rates during recent and past exposure, with the highest incidence during non-use; particularly observed for ischemic stroke. Previously reported incidence rates of stroke are in the same range as in our study but did not include detailed information on VKA use from newly diagnosed AF patients of the recent years [13, 14]. For ischemic stroke, the incidence rate reported by Go *et al*. [15] in patients taking warfarin and for patients not taking warfarin is in agreement with our study. For MI, the incidence rate was comparable between current and recent VKA use, whereas past and non-use showed a trend to higher incidence in this study. It has been reported that AF patients are at higher risk for coronary events [16]. Sensitivity analysis for cardiovascular events based on diagnosis recorded by the general practitioner showed a similar trend with regard to the incidences stratified by periods of VKA use, but resulted in slightly higher incidences than those based on hospital data for MI and stroke. This might be explained as the analysis of GP records is more likely to include records of patients for non-hospitalised events, and MIs leading to death without hospitalisation.

The incidence of bleeding events in this study was comparable during current and recent VKA use whereas lower incidences were shown during past and non-use, which might be explained by the nature of VKA treatment. Go *et al*. [15] reports incidence rates of intracranial haemorrhage based on hospitalisations of 0.5 (95% CI 0.4 to 0.6) and 0.2 (0.2 to 0.3)/100 person-years (PY) among patients taking and not taking warfarin. This compares to 0.4 (0.3 to 0.5) and 0.6 (0.4 to 1.1)/100 PY during current and recent exposure, and 0.2 (0.1 to 0.4) and 0.4 (0.3 to 0.5)/100 PY for past and non-use in this study, respectively. The similar intracranial bleeding incidence during non VKA use compared to current exposure might be explained by contraindications associated with higher risk for intracranial bleedings. For gastrointestinal bleedings, Go *et al*. [15] did not find differences between patients taking and not taking warfarin; this study showed a trend towards higher incidences during recent VKA exposure compared to current exposure, 1.0 (0.7 to 1.4) for past and 1.1 (1.0 to 1.3) per 100 PY for non-use. Thus, the bleeding incidence results of

our study indicate a potentially increased risk for AF patients that recently stopped VKA treatment compared to periods of past exposure and non-use. However, these findings could be related to a general worsening of health status or to the fact that these patients acquired additional bleeding risk factors (e.g. start of aspirin use), which lead to stopping of VKA treatment. We used a comprehensive list of 96 ICD-10 codes (Table S3, supporting information) to define bleeding events including additional sites (e.g., kidney, uterus). Compared to the code lists published by Hansen and Olesen, which we used for sensitivity analyses, the bleeding rates in our main analysis were higher than the rates observed in the sensitivity analyses (Table S2, supporting information) [11,12]. However, the rates in our sensitivity analyses correspond well to the published data [11]. The proportion of fatal bleeding events was highest among patients suffering from intracranial bleedings (38% overall, compared to 3% among extracranial and 6% among GI bleedings). The proportion of fatal intracranial bleeding events was higher among current (50%), recent (40%) and past (63%) VKA exposure compared to non-use (22%), however the numbers for recent and past exposures were small. This is in line with the ATRIA cohort study conducted between 1996 and 1997, which reported that the risk of death within 30 days after hospitalisation for warfarin-associated intracranial bleedings was 48.6% [17]. However, there was neither stratification on VKA exposure, or on bleeding site. While in this study any-cause hospitalisation was lowest during current VKA use, there was no clear pattern for AF-hospitalisations dependent on VKA treatment with lowest incidences during past VKA exposure. Mortality was substantially lower during current VKA exposure compared to other exposure periods. This might be related to the protective effect of VKAs for stroke and to the healthy user effect, i.e. patients currently taking VKAs might be more adherent to medication and more health conscious.

The strengths of the study include the large sample size of a population representative of the UK population. Although only a proportion of the eligible population could be linked to the HES/ONS data, the comparison of patient characteristics between the overall cohort of incident AF patients and the final study population showed no major differences (Table S5, supporting information) and thus should remain representative for the incident AF population in the UK. In addition, a recent comparison of patients included or not in the linkage found that there were no differences in patient characteristics [18]. Linkage of data from hospital and official death certificates added valuable information and confirmed data validity. Another important strength is the up-to-date data. The timing of the data collection (2005-2010) allows for the most appropriate understanding of event rates compared to older data sets, for example, from the 1990s, especially in the context of changing concomitant treatments of comorbidities (e.g. intensity of use of antihypertensives, lipid lowering drugs and newer antiarrhythmic drugs) and immediately before NOACs were launched [19]. This data will help to set future data on NOACs into perspective.

With the introduction of NOACs into the market a new era has commenced, that will change the disease and treatment landscape and its' impact on routine clinical practice needs to be established using such studies. In addition, incident cohorts are the most valid basis for incidence estimates [20]. For appropriate thromboprophylaxis it is crucial to assess stroke and bleeding risk using current risk stratification scores [21]. In addition, information of our study about current incidence rates of common comorbidities and outcomes in the AF patient population will help to guide clinical management decisions of AF patients and put further data into perspective.

Nonetheless, there are limitations to this study. We did not have information on the diagnostic criteria used for the AF diagnosis, such as electrocardiograms, and underdiagnosis of AF is a current public health problem. However, a previous CPRD study - although inclusion criteria of AF patients differ slightly compared to our study - reported a high level of validity in the recording of AF by GPs [22]. The procedures used to diagnose strokes and clinical presentation of stroke cases are not recorded in CPRD. In our main analysis, only hospitalised outcomes are included, likely leading to lower incidence rates as less severe outcomes will not lead to hospitalisation (e.g. minor bleedings or transient ischemic attack). This is also shown in our sensitivity analysis of cardiovascular outcomes. Furthermore, drug exposure was based on prescriptions and/or INR measurements and assumptions on dosages had to be made. We cannot exclude that patients may be supplied with VKAs for more than a month's needs, and that some patients currently taking VKAs were misclassified under recent exposure. No over-the-counter medications such as aspirin could be considered. Our analyses on haemorrhagic stroke and intracranial haemorrhage are based on low numbers of events and no firm conclusions can be drawn. Only crude estimates were estimated which do not allow any assessment of causal relationships between exposures and outcomes. Finally, this is an observational study and patients were not randomised to a particular treatment. Changes in the exposure to VKAs are expected to be influenced by confounding factors. Stopping of preventive treatments like VKA treatment can occur in patients with worsening of the health status [23-25]. These patients would provide data to the 'recent exposure' category and thus might increase the incidence of outcome events.

Direct comparisons of incidence rates between CPRD data and randomised controlled trials (RCTs) have limitations due to differences in inclusion/exclusion criteria and outcome definitions and adjudication. However, for outcomes that are more consistently defined (such as mortality), a comparison of incidence rates in RCTs and actual clinical practice may show the external validity and representativeness of the RCT evidence. It has been proposed that mortality rates in RCTs provide an important guide to the representativeness of a RCT [26]. In such RCTs mortality rates were mainly between 3 and 5/100 PY [3, 27, 28]. In this CPRD study the mortality rate was well within this range for current VKA use, but higher for

recent, past and non-use. Our study shows major differences when stratifying patient groups into current, recent, past and no VKA use. Patients that recently stopped VKA treatment show the highest mortality rate, which might be related to the circumstance that effective drugs are likely not used by patients near death, due to selective non-prescription by physicians or non-adherence by patients [23].

In conclusion, this study provides a comprehensive description of incident AF patients and estimates of the incidence of vascular and other outcomes in patients with AF in a GPs' database linked to hospital data from the years immediately before new oral anticoagulants were introduced.

Strengths and limitations of this study

- Stratified analysis of current, recent and past exposure, as well as periods of no vitamin K antagonist (VKA) use, showing substantial differences in comorbidities and outcome events.

- Analysis of only newly diagnosed patients with atrial fibrillation (AF), reducing the risk of bias and reflecting the most current population diagnosed with AF.

- Patients were not randomised to a particular treatment.

- Information on the diagnostic criteria used for the diagnosis of AF is not available in the Clinical Practice Research Datalink database.

- Only crude estimates were estimated which do not allow assessment of causal associations.

References

1. European Heart Rhythm Association; European Association for Cardio-Thoracic Surgery, Camm AJ, Kirchhof P, Lip GY, Schotten U, Savelieva I, Ernst S, Van Gelder IC, Al-Attar N, Hindricks G, Prendergast B, Heidbuchel H, Alfieri O, Angelini A, Atar D, Colonna P, De Caterina R, De Sutter J, Goette A, Gorenek B, Heldal M, Hohloser SH, Kolh P, Le Heuzey JY, Ponikowski P, Rutten FH. Guidelines for the management of atrial fibrillation: the Task Force for the Management of atrial Fibrillation of the European Society of Cardiology (ESC). Eur Heart J. 2010;31(19):2369-429.
2. Connolly SJ, Pogue J, Eikelboom J, Flaker G, Commerford P, Franzosi MG, Healey JS, Yusuf S; ACTIVE W Investigators. Benefit of oral anticoagulant over antiplatelet therapy in atrial fibrillation depends on the quality of international normalized ratio control achieved by centers and countries as measured by time in therapeutic range. Circulation. 2008;118(20):2029-37.
3. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, Pogue J, Reilly PA, Themeles E, Varrone J, Wang S, Alings M, Xavier D, Zhu J, Diaz R, Lewis BS, Darius H, Diener HC, Joyner CD, Wallentin L; RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med. 2009;361(12):1139-51.
4. Jabre P, Jouven X, Adnet F, Thabut G, Bielinski SJ, Weston SA, Roger VL. Atrial fibrillation and death after myocardial infarction: a community study. Circulation. 2011;123(19):2094-100.
5. Williams T, van staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. Ther Adv Drug Saf. 2012;3(2):89-99.
6. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69(1):4-14.
7. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA. 2001;285(22):2864-70.
8. Lane DA, Lip GYH. Use of the CHA2DS2-VASc and HAS-BLED scores to aid decision making for thromboprophylaxis in nonvalvular atrial fibrillation. Circulation 2012;126:860-5.
9. Rietbrock S, Heeley E, Plumb J, van Staa T. Chronic atrial fibrillation: Incidence, prevalence, and prediction of stroke using the Congestive heart failure, Hypertension, Age >75, Diabetes mellitus, and prior Stroke or transient ischemic attack (CHADS2) risk stratification scheme. Am Heart J. 2008;156(1):57-64.
10. Lewis JD, Brensinger C. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. Pharmacoepidemiol Drug Saf. 2004;13(7):437-41.
11. Hansen ML, Sørensen R, Clausen MT, Fog-Petersen ML, Raunsø J, Gadsbøll N, Gislason GH, Folke F, Andersen SS, Schramm TK, Abildstrøm SZ, Poulsen HE, Køber L, Torp-Pedersen C. Risk of bleeding with single, dual, or triple therapy with warfarin, aspirin, and clopidogrel in patients with atrial fibrillation. Arch Intern Med. 2010;170(16):1433-41.
12. Olesen JB, Lip GY, Lindhardsen J, Lane DA, Ahlehoff O, Hansen ML, Raunsø J, Tolstrup JS, Hansen PR, Gislason GH, Torp-Pedersen C. Risks of thromboembolism and bleeding with thromboprophylaxis in patients with atrial fibrillation: A net clinical benefit analysis using a 'real world' nationwide cohort study. Thromb Haemost. 2011;106(4):739-49.

13. Frost L, Vukelic Andersen L, Vestergaard P, Husted S, Mortensen LS. Trends in risk of stroke in patients with a hospital diagnosis of nonvalvular atrial fibrillation: National Cohort Study in Denmark, 1980-2002. Neuroepidemiology. 2006;26(4):212-9.
14. Friberg L, Hammar N, Rosenqvist M. Stroke in paroxysmal atrial fibrillation: report from the Stockholm Cohort of Atrial Fibrillation. Eur Heart J 2010;31:967-75.
15. Go AS, Hylek EM, Chang Y, Phillips KA, Henault LE, Capra AM, Jensvold NG, Selby JV, Singer DE. Anticoagulation therapy for stroke prevention in atrial fibrillation: how well do randomized trials translate into clinical practice? JAMA. 2003;290(20):2685-92.
16. Ruigómez A, Johansson S, Wallander MA, Edvardsson N, García Rodríguez LA. Risk of cardiovascular and cerebrovascular events after atrial fibrillation diagnosis. Int J Cardiol. 2009;136(2):186-92.
17. Fang MC, Go AS, Chang Y, Hylek EM, Henault LE, Jensvold NG, Singer DE. Death and disability from warfarin-associated intracranial and extracranial hemorrhages. Am J Med. 2007;120(8):700-5.
18. Gallagher AM, Puri S, van Staa TP. Linkage of the General Practice Research Database (GPRD) with Other Data Sources. Pharmacoepidemiol Drug Saf. 2011;20:S230.
19. Parkinson J, Davis S, van Staa TP. The General Practice Research (GPRD) Database: Now and the Future. In Pharmacovigilance Editor R Mann. Second edition. John Wiley & Sons. 2007:341-8.
20. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19:858-68.
21. Lip GYH. Stroke and bleeding risk assessment in atrial fibrillation: when, how, and why? Eur Heart J. 2013;34:1041-9.
22. Ruigómez A, Johansson S, Wallander M-A, Rodriguez LA. Incidence of chronic atrial fibrillation in general practice and its treatment pattern. Journal of Clinical Epidemiology. 2002;55:358-63.
23. Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology. 2001;12(6):682-9.
24. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. N Engl J Med 1998;338:1516-20.
25. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution--a simulation study. Am J Epidemiol. 2010;172:843-54.
26. Silberman G, Droitcour JA, Scullin EW. Cross design synthesis: a new strategy for medical effectiveness research. Darby, USA: Diane Publishing Co, 1992.
27. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, Breithardt G, Halperin JL, Hankey GJ, Piccini JP, Becker RC, Nessel CC, Paolini JF, Berkowitz SD, Fox KA, Califf RM; ROCKET AF Investigators. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med. 2011;365(10):883-91.
28. Granger CB, Alexander JH, McMurray JJ, Lopes RD, Hylek EM, Hanna M, Al-Khalidi HR, Ansell J, Atar D, Avezum A, Bahit MC, Diaz R, Easton JD, Ezekowitz JA, Flaker G, Garcia D, Geraldes M, Gersh BJ, Golitsyn S, Goto S, Hermosillo AG, Hohnloser SH, Horowitz J, Mohan P, Jansky P, Lewis BS, Lopez-Sendon JL, Pais P, Parkhomenko A, Verheugt FW, Zhu J, Wallentin L; ARISTOTLE Committees and Investigators. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med. 2011;365(11):981-92.

Supporting information

Details on HES:

The national registry of hospital admissions (HES) data contain details of all admissions to National Health Service (NHS) hospitals in England. The patients include private patients and those treated in NHS hospitals residing outside of England, as well as care delivered by treatment centres (including the independent sector) funded by the NHS. All NHS trusts in England, including acute hospitals, primary care trusts and mental health trusts are included. HES data are administrative data, designed to enable secondary use for healthcare analyses. The data are collected during a patient's time at hospital and are used to allow hospitals to be paid for the care they deliver. As the data are extracted from patients' notes, they may also reflect the quality of clinical record keeping. For each hospitalised patient, the hospital charts are reviewed, dates of admission and discharge and main diagnoses are extracted, coded by coding staff and collated nationally into HES.

Table S1: Incidence of vascular outcomes overall and in relation to VKA use for the cohort of AF patients eligible for linkage, n=16,513

	Overall			Current Use			Recent Use			Past Use			No use		
Event	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**
Vascular	1436	29854	4.8 (4.6 to 5.1)	272	10776	2.5 (2.2 to 2.8)	78	2183	3.6 (2.8 to 4.5)	151	3577	4.2 (3.6 to 5.0)	935	13319	7.0 (6.6 to 7.5)
Fatal	111	29854	0.4 (0.3 to 0.4)	23	10776	0.2 (0.1 to 0.3)	15	2183	0.7 (0.4 to 1.1)	9	3577	0.3 (0.1 to 0.5)	64	13319	0.5 (0.4 to 0.6)
Non-fatal	1325	29854	4.4 (4.2 to 4.7)	249	10776	2.3 (2.0 to 2.6)	63	2183	2.9 (2.2 to 3.7)	142	3577	4.0 (3.3 to 4.7)	871	13319	6.5 (6.1 to 7.0)
MI	518	31176	1.7 (1.5 to 1.8)	101	11419	0.9 (0.7 to 1.1)	19	2300	0.8 (0.5 to 1.3)	50	3723	1.3 (1.0 to 1.8)	348	13734	2.5 (2.3 to 2.8)
Fatal	57	31176	0.2 (0.1 to 0.2)	12	11419	0.1 (0.1 to 0.2)	5	2300	0.2 (0.1 to 0.5)	2	3723	†	38	13734	0.3 (0.2 to 0.4)
Non-fatal	461	31176	1.5 (1.3 to 1.6)	89	11419	0.8 (0.6 to 1.0)	14	2300	0.6 (0.3 to 1.0)	48	3723	1.3 (1.0 to 1.7)	310	13734	2.3 (2.0 to 2.5)
Stroke	937	30599	3.1 (2.9 to 3.3)	173	11025	1.6 (1.3 to 1.8)	62	2243	2.8 (2.1 to 3.5)	102	3675	2.8 (2.3 to 3.4)	600	13657	4.4 (4.0 to 4.8)
Haemorrhagic	62	31884	0.2 (0.1 to 0.2)	19	11643	0.2 (0.1 to 0.3)	7	2353	0.3 (0.1 to 0.6)	5	3821	†	31	14067	0.2 (0.1 to 0.3)
Ischemic	682	30936	2.2 (2.0 to 2.4)	128	11163	1.1 (1.0 to 1.4)	43	2267	1.9 (1.4 to 2.6)	81	3717	2.2 (1.7 to 2.7)	430	13789	3.1 (2.8 to 3.4)
Unspecified	241	31603	0.8 (0.7 to 0.9)	39	11522	0.3 (0.2 to 0.5)	17	2333	0.7 (0.4 to 1.2)	19	3788	0.5 (0.3 to 0.8)	166	13959	1.2 (1.0 to 1.4)

Table S1 (cont.): Incidence of vascular outcomes overall and in relation to VKA use for the cohort of AF patients eligible for linkage, n=16,513

	Overall			Current Use			Recent Use			Past Use			No use		
Event	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**
Fatal Stroke	57	30599	0.2 (0.1 to 0.2)	13	11025	0.1 (0.1 to 0.2)	10	2243	0.4 (0.2 to 0.8)	8	3675	0.2 (0.1 to 0.4)	26	13657	0.2 (0.1 to 0.3)
Haemorrhagic	15	31884	†	9	11643	†	1	2353	†	2	3821	†	3	14067	†
Ischemic	22	30936	†	3	11163	†	3	2267	†	3	3717	†	13	13789	0.1 (0.1 to 0.2)
Unspecified	25	31603	0.1 (0.1 to 0.1)	5	11522	†	6	2333	0.3 (0.1 to 0.6)	3	3788	†	11	13959	†
Non-fatal Stroke	880	30599	2.9 (2.7 to 3.1)	160	11025	1.5 (1.2 to 1.7)	52	2243	2.3 (1.7 to 3.0)	94	3675	2.6 (2.1 to 3.1)	574	13657	4.2 (3.9 to 4.6)
Haemorrhagic	47	31884	0.1 (0.1 to 0.2)	10	11643	†	6	2353	0.3 (0.1 to 0.6)	3	3821	†	28	14067	0.2 (0.1 to 0.3)
Ischemic	660	30936	2.1 (2.0 to 2.3)	125	11163	1.1 (0.9 to 1.3)	40	2267	1.8 (1.3 to 2.4)	78	3717	2.1 (1.7 to 2.6)	417	13789	3.0 (2.7 to 3.3)
Unspecified	216	31603	0.7 (0.6 to 0.8)	34	11522	0.3 (0.2 to 0.4)	11	2333	0.5 (0.3 to 0.8)	16	3788	0.4 (0.2 to 0.7)	155	13959	1.1 (0.9 to 1.3)
SPAE	29	31897	0.1 (0.1 to 0.1)	11	11633	†	0	2354	†	4	3828	†	14	14082	0.1 (0.1 to 0.2)
Fatal	1	31897	†	0	11633	†	0	2354	†	1	3828	†	0	14082	†
Non-fatal	28	31897	0.1 (0.1 to 0.1)	11	11633	†	0	2354	†	3	3828	†	14	14082	0.1 (0.1 to 0.2)

* Number of cases per 100 person years

† Not enough events to calculate

Note: Outcomes are based on Read codes from primary care

AF: atrial fibrillation; MI: myocardial infarction; N: number of outcome events; PY: person-years; SPAE: systemic pulmonary arterial embolism; VKA: vitamin K antagonist.

Table S2: Incidence of bleeding outcomes overall and in relation to VKA use for the cohort of AF patients eligible for linkage, n=16,513

	Overall			Current Use			Recent Use			Past Use			No use		
Event	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**	**N**	**PY**	**Incidence* (95% CI)**
All Bleed	782	31004	2.5 (2.3 to 2.7)	335	11340	3.0 (2.6 to 3.3)	77	2262	3.4 (2.7 to 4.3)	71	3637	2.0 (1.5 to 2.5)	299	13765	2.2 (1.9 to 2.4)
Intracranial	120	31854	0.4 (0.3 to 0.5)	46	11654	0.4 (0.3 to 0.5)	15	2353	0.6 (0.4 to 1.1)	8	3811	0.2 (0.1 to 0.4)	51	14036	0.4 (0.3 to 0.5)
Extra-cranial	662	31090	2.1 (2.0 to 2.3)	289	11349	2.5 (2.3 to 2.9)	62	2268	2.7 (2.1 to 3.5)	63	3657	1.7 (1.3 to 2.2)	248	13816	1.8 (1.6 to 2.0)
Gastrointestinal	209	31703	0.7 (0.6 to 0.8)	64	11609	0.6 (0.4 to 0.7)	18	2342	0.8 (0.5 to 1.2)	23	3768	0.6 (0.4 to 0.9)	104	13984	0.7 (0.6 to 0.9)
Fatal Bleed	64	31004	0.2 (0.2 to 0.3)	27	11340	0.2 (0.2 to 0.3)	9	2262	0.4 (0.2 to 0.8)	6	3637	0.2 (0.1 to 0.4)	22	13765	0.2 (0.1 to 0.2)
Intracranial	45	31854	0.1 (0.1 to 0.2)	23	11654	0.2 (0.1 to 0.3)	6	2353	0.3 (0.1 to 0.6)	5	3811	0.1 (0.0 to 0.3)	11	14036	0.1 (0.0 to 0.1)
Extra-cranial	19	31090	0.1 (0.0 to 0.1)	4	11349	0.0 (0.0 to 0.1)	3	2268	0.1 (0.0 to 0.4)	1	3657	0.0 (0.0 to 0.2)	11	13816	0.1 (0.0 to 0.1)
Gastrointestinal	17	31703	0.1 (0.0 to 0.1)	3	11609	0.0 (0.0 to 0.1)	3	2342	0.1 (0.0 to 0.4)	1	3768	0.0 (0.0 to 0.1)	10	13984	0.1 (0.0 to 0.1)
Non-Fatal Bleed	718	31004	2.3 (2.1 to 2.5)	308	11340	2.7 (2.4 to 3.0)	68	2262	3.0 (2.3 to 3.8)	65	3637	1.8 (1.4 to 2.3)	277	13765	2.0 (1.8 to 2.3)
Intracranial	75	31854	0.2 (0.2 to 0.3)	23	11654	0.2 (0.1 to 0.3)	9	2353	0.4 (0.2 to 0.7)	3	3811	0.1 (0.0 to 0.2)	40	14036	0.3 (0.2 to 0.4)
Extra-cranial	643	31090	2.1 (1.9 to 2.2)	285	11349	2.5 (2.2 to 2.8)	59	2268	2.6 (2.0 to 3.4)	62	3657	1.7 (1.3 to 2.2)	237	13816	1.7 (1.5 to 1.9)
Gastrointestinal	192	31703	0.6 (0.5 to 0.7)	61	11609	0.5 (0.4 to 0.7)	15	2342	0.6 (0.4 to 1.1)	22	3768	0.6 (0.4 to 0.9)	94	13984	0.7 (0.5 to 0.8)

* Number of cases per 100 person years

Note: Outcomes are based on previously published ICD-10 codes (Hansen et al, 2010; Olesen et al, 2011)

AF: atrial fibrillation; GI: gastrointestinal; ICD-10: International Classification of Disease, version 10; PY: person-years; VKA: vitamin K antagonists.

Table S3: List of ICD-10 codes of bleeding definition

ICD-10 code	ICD-10 Term	Type of bleed	included in literature definition
I85.0	Oesophageal varices with bleeding	Gastrointestinal	-
K25.0	Gastric ulcer, acute with haemorrhage	Gastrointestinal	+
K25.2	Gastric ulcer, acute with both haemorrhage and perforation	Gastrointestinal	-
K25.4	Gastric ulcer, chronic or unspecified with haemorrhage	Gastrointestinal	+
K25.6	Chronic or unspecified with both haemorrhage and perforation	Gastrointestinal	-
K26.0	Duodenal ulcer, acute with haemorrhage	Gastrointestinal	+
K26.2	Duodenal ulcer, acute with both haemorrhage and perforation	Gastrointestinal	-
K26.4	Duodenal ulcer, chronic or unspecified with haemorrhage	Gastrointestinal	+
K26.6	Chronic or unspecified with both haemorrhage and perforation	Gastrointestinal	-
K27.0	Peptic ulcer, acute with haemorrhage	Gastrointestinal	+
K27.2	Peptic ulcer, acute with both haemorrhage and perforation	Gastrointestinal	-
K27.4	Peptic ulcer, chronic or unspecified with haemorrhage	Gastrointestinal	-
K27.6	Chronic or unspecified with both haemorrhage and perforation	Gastrointestinal	-
K28.0	Gastrojejunal ulcer, acute with haemorrhage	Gastrointestinal	+
K28.2	Acute with both haemorrhage and perforation	Gastrointestinal	-
K28.4	Gastrojejunal ulcer, chronic or unspecified with haemorrhage	Gastrointestinal	-
K28.6	Chronic or unspecified with both haemorrhage and perforation	Gastrointestinal	-
K29.0	Acute haemorrhagic gastritis	Gastrointestinal	-
K62.5	Haemorrhage of anus and rectum	Gastrointestinal	-
K66.1	Haemoperitoneum	Gastrointestinal	-
K92.0	Haematemesis	Gastrointestinal	+
K92.1	Melaena	Gastrointestinal	+
K92.2	Gastrointestinal haemorrhage, unspecified	Gastrointestinal	+

Table S3 (cont.): List of International Classification of Disease, version 10 (ICD-10) codes of bleeding definition

ICD-10 code	ICD-10 Term	Type of bleed	included in literature definition
I60	Subarachnoid haemorrhage	Intracranial	+
I60.0	Subarachnoid haemorrhage from carotid siphon and bifurcation	Intracranial	+
I60.1	Subarachnoid haemorrhage from middle cerebral artery	Intracranial	+
I60.2	Subarachnoid haemorrhage from anterior communicating artery	Intracranial	+
I60.3	Subarachnoid haemorrhage from posterior communicating artery	Intracranial	+
I60.4	Subarachnoid haemorrhage from basilar artery	Intracranial	+
I60.5	Subarachnoid haemorrhage from vertebral artery	Intracranial	+
I60.6	Subarachnoid haemorrhage from other intracranial arteries	Intracranial	+
I60.7	Subarachnoid haemorrhage from intracranial artery, unspec	Intracranial	+
I60.8	Other subarachnoid haemorrhage	Intracranial	+
I60.9	Subarachnoid haemorrhage, unspecified	Intracranial	+
I61	Intracerebral haemorrhage	Intracranial	+
I61.0	Intracerebral haemorrhage in hemisphere, subcortical	Intracranial	+
I61.1	Intracerebral haemorrhage in hemisphere, cortical	Intracranial	+
I61.2	Intracerebral haemorrhage in hemisphere, unspecified	Intracranial	+
I61.3	Intracerebral haemorrhage in brain stem	Intracranial	+
I61.4	Intracerebral haemorrhage in cerebellum	Intracranial	+
I61.5	Intracerebral haemorrhage, intraventricular	Intracranial	+
I61.6	Intracerebral haemorrhage, multiple localized	Intracranial	+
I61.8	Other intracerebral haemorrhage	Intracranial	+
I61.9	Intracerebral haemorrhage, unspecified	Intracranial	+
I62	Other nontraumatic intracranial haemorrhage	Intracranial	+
I62.0	Subdural haemorrhage (acute)(nontraumatic)	Intracranial	+
I62.1	Nontraumatic extradural haemorrhage	Intracranial	+
I62.9	Intracranial haemorrhage (nontraumatic), unspecified	Intracranial	+
I69.0	Sequelae of subarachnoid haemorrhage	Intracranial	+
I69.1	Sequelae of intracerebral haemorrhage	Intracranial	+
I69.2	Sequelae of other nontraumatic intracranial haemorrhage	Intracranial	+
S06.4	Epidural haemorrhage	Intracranial	+
S06.5	Traumatic subdural haemorrhage	Intracranial	+
S06.6	Traumatic subarachnoid haemorrhage	Intracranial	+

Table S3 (cont.): List of ICD-10 codes of bleeding definition

ICD-10 code	ICD-10 Term	Type of bleed	included in literature definition
J94.2	Haemothorax	Airway	+
R04	Haemorrhage from respiratory passages	Airway	+
R04.0	Epistaxis	Airway	+
R04.1	Haemorrhage from throat	Airway	+
R04.2	Haemoptysis	Airway	+
R04.8	Haemorrhage from other sites in respiratory passages	Airway	+
R04.9	Haemorrhage from respiratory passages, unspecified	Airway	+
N02	Recurrent and persistent haematuria	Urinary	+
N02.6	Recurrent and persistent haematuria, dense deposit disease	Urinary	+
N02.8	Recurrent and persistent haematuria, other	Urinary	+
N02.9	Recurrent and persistent haematuria, unspecified	Urinary	+
R31	Unspecified haematuria	Urinary	+
D68.3	Haemorrhagic disorder due to circulating anticoagulants	Blood	-
H35.6	Retinal haemorrhage	Eyes	-
H43.1	Vitreous haemorrhage	Eyes	-
H45.0	Vitreous haemorrhage in diseases classified elsewhere	Eyes	-
H92.2	Otorrhagia	Ear	-
I23.0	Haemopericardium as curr comp folow acut myocard infarct	Heart	-
I31.2	Haemopericardium, not elsewhere classified	Heart	-
I71.1	Thoracic aortic aneurysm, ruptured	Thorax	-
I71.3	Abdominal aortic aneurysm, ruptured	Abdominal	-
I71.5	Thoracoabdominal aortic aneurysm, ruptured	Abdominal	-
I84.1	Internal haemorrhoids with other complications	Rectal	-
I84.3	External thrombosed haemorrhoids	Rectal	-
I84.4	External haemorrhoids with other complications	Rectal	-
I84.8	Unspecified haemorrhoids with other complications	Rectal	-
N42.1	Congestion and haemorrhage of prostate	Renal	-
N83.6	Haematosalpinx	Renal	-
N85.7	Haematometra	Renal	-
N89.7	Haematocolpos	Renal	-
N92.1	Excessive and frequent menstruation with irregular cycle	Uterine	-
N92.5	Other specified irregular menstruation	Uterine	-
N92.6	Irregular menstruation, unspecified	Uterine	-
N93	Other abnormal uterine and vaginal bleeding	Uterine	-
N93.8	Other specified abnormal uterine and vaginal bleeding	Uterine	-
N93.9	Abnormal uterine and vaginal bleeding, unspecified	Uterine	-
N95.0	Postmenopausal bleeding	Uterine	-
R23.3	Spontaneous ecchymoses	Respiratory	-
R58	Haemorrhage, not elsewhere classified	Respiratory	-
I71.8	Aortic aneurysm of unspecified site, ruptured	Unknown	-
T79.2	Traumatic secondary and recurrent haemorrhage	Unknown	-
T81.0	Haemorrhage and haematoma complicating a procedure NEC	Unknown	-

Table S4: Read codes for the definition of atrial fibrillation (AF)

Read code	Read term
3272.00	ECG: atrial fibrillation
662S.00	Atrial fibrillation monitoring
6A9..00	Atrial fibrillation annual review
7936A00	Implant intravenous pacemaker for atrial fibrillation
9Os..00	Atrial fibrillation monitoring administration
9Os0.00	Atrial fibrillation monitoring first letter
9Os1.00	Atrial fibrillation monitoring second letter
9Os2.00	Atrial fibrillation monitoring third letter
9Os3.00	Atrial fibrillation monitoring verbal invite
9Os4.00	Atrial fibrillation monitoring telephone invite
G573.00	Atrial fibrillation and flutter
G573000	Atrial fibrillation
G573200	Paroxysmal atrial fibrillation
G573300	Non-rheumatic atrial fibrillation
G573400	Permanent atrial fibrillation
G573500	Persistent atrial fibrillation
G573z00	Atrial fibrillation and flutter NOS

Table S5: Comparison between total AF population in CPRD and the AF population eligible for HES/ONS linkage

Characteristics	All incident AF patients N=42,008	AF patients eligible for HES/ONS linkage N=16,513
Length of follow-up (years):		
mean (sd)	2.4 (1.6)	1.9 (1.4)
median (Range)	2.2 (0 - 6)	1.7 (0 - 5)
Age:		
mean (sd)	74 (12)	74 (12.0)
median (Range)	76 (18 - 107)	76 (21 - 107)
<40	502 (1.2%)	192 (1.2%)
40 - <50	1,217 (2.9%)	462 (2.8%)
50 - <65	6,609 (15.7%)	2,549 (15.4%)
65 - <75	10,656 (25.4%)	4,120 (25.0%)
75 - <80	7,458 (17.8%)	2,913 (17.6%)
>=80	15,566 (37.1%)	6,277 (38.0%)
Female	19,934 (47.5%)	7,921 (48.0%)
Male	22,074 (52.5%)	8,592 (52.0%)
Year of Diagnosis		
2005	8,065 (19.2%)	3,443 (20.9%)
2006	8,454 (20.1%)	3,516 (21.3%)
2007	8,075 (19.2%)	3,428 (20.8%)
2008	8,125 (19.3%)	3,548 (21.5%)
2009	8,027 (19.1%)	2,578 (15.6%)
Body Mass Index (kg/m2):		
mean (SD)	28 (6)	28 (6)
median (range)	27 (12 - 70)	27 (12 - 70)
<25	13,216 (31.5%)	5,312 (32.2%)
25 - <30	14,237 (33.9%)	5,611 (34.0%)
30 - <35	7,306 (17.4%)	2,821 (17.1%)
>=35	3,950 (9.4%)	1,532 (9.3%)
BMI: Unknown	3,299 (7.9%)	1,237 (7.5%)
Smoking Status		
Non Smoker	17,391 (41.4%)	6,822 (41.3%)
Ex Smoker	18,694 (44.5%)	7,470 (45.2%)
Smoker	5,235 (12.5%)	2,003 (12.1%)
Smoking Status: Unknown	688 (1.6%)	218 (1.3%)
Drinking Status		
Non Drinker	4,713 (11.2%)	1,692 (10.2%)
Ex Drinker	4,251 (10.1%)	1,623 (9.8%)
Drinker	28,749 (68.4%)	11,601 (70.3%)
Drinking Status: Unknown	4,295 (10.2%)	1,597 (9.7%)

Table S5 (cont.): Comparison between total AF population in CPRD and the AF population eligible for HES/ONS linkage

Characteristics	All incident AF patients N=42,008	AF patients eligible for HES/ONS linkage N=16,513
Deprivation quintile, based on patient's home		
0 (least deprived)	3,342 (8.0%)	2,753 (16.7%)
1	3,403 (8.1%)	2,896 (17.5%)
2	3,133 (7.5%)	2,635 (16.0%)
3	2,931 (7.0%)	2,451 (14.8%)
4 (most deprived)	2,653 (6.3%)	2,226 (13.5%)
Deprivation level unknown	26,546 (63.2%)	3,552 (21.5%)
$CHADS_2$ risk score:		
mean (SD)	1.48 (1.18)	1.50 (1.18)
median (IQR)	1 (1-2)	1 (1-2)
0	9,086 (21.6%)	3,493 (21.2%)
1	14,080 (33.5%)	5,508 (33.4%)
2	11,544 (27.5%)	4,605 (27.9%)
3	4,498 (10.7%)	1,778 (10.8%)
4	2,157 (5.1%)	846 (5.1%)
5	588 (1.4%)	260 (1.6%)
6	55 (0.1%)	23 (0.1%)
CHA_2DS_2-VASc risk score:		
mean (SD)	3.01 (1.76)	3.04 (1.77)
median (IQR)	3 (2-4)	3 (2-4)
0	3,294 (7.8%)	1,258 (7.6%)
1	5,195 (12.4%)	1,974 (12.0%)
2	7,978 (19.0%)	3,099 (18.8%)
3	10,244 (24.4%)	4,049 (24.5%)
4	7,996 (19.0%)	3,234 (19.6%)
5	3,330 (7.9%)	1,290 (7.8%)
6	2,203 (5.2%)	861 (5.2%)
7	1,389 (3.3%)	587 (3.6%)
8	353 (0.8%)	153 (0.9%)
9	26 (0.1%)	8 (0.0%)

Table S5 (cont.): Comparison between total AF population in CPRD and the AF population eligible for HES/ONS linkage

Characteristics	All incident AF patients N=42,008	AF patients eligible for HES/ONS linkage N=16,513
History of*:		
Acute myocardial infarction	4.089 (9.7%)	1,621 (9.8%)
Stroke	4,519 (10.8%)	1,844 (11.2%)
Systemic pulmonary arterial embolism	115 (0.3%)	49 (0.3%)
Vascular event	8,098 (19.3%)	3,273 (19.8%)
Coronary artery disease	7,201 (17.1%)	2,776 (16.8%)
Congestive heart failure	3,415 (8.1%)	1,357 (8.2%)
Hypertension	20,458 (48.7%)	8,113 (49.1%)
Diabetes	6,625 (15.8%)	2,526 (15.3%)
Liver Impairment	808 (1.9%)	337 (2.0%)
Renal Impairment	6,531 (15.5%)	2,513 (15.2%)
Treatment with**:		
Anti-arrhythmics	3,372 (8.0%)	1,422 (8.6%)
Statins	14,908 (35.5%)	5,722 (34.7%)
Antiplatelets	20,935 (49.8%)	8,148 (49.3%)
Aspirin	19,519 (46.5%)	7,599 (46.0%)
Low dose aspirin (<100mg)	18,841 (44.9%)	7,331 (44.4%)
High dose aspirin (100mg+)	881 (2.1%)	346 (2.1%)
Clopidogrel	2,130 (5.1%)	816 (4.9%)
Other anti-platelets	884 (2.1%)	309 (1.9%)
Anticoagulants	4,796 (11.4%)	1,816 (11.0%)
Anti-diabetics	3,876 (9.2%)	1,498 (9.1%)
Oral anti-diabetics	3,374 (8.0%)	1,301 (7.9%)
Insulin	937 (2.2%)	365 (2.2%)
Other injectable anti-diabetics	6 (<0.1%)	2 (<0.1%)
Antihypertensives***	27,250 (64.9%)	10,747 (65.1%)

* Comorbidities were measured prior to the index date in GPRD.
** Treatments were measured in the 90 days prior to index date.
*** Antihypertensives include Angiotensin-converting enzyme inhibitors, Angiotensin-II receptor antagonists, Beta-adrenoceptor blocking drugs, Calcium channel blockers, Centrally-acting antihypertensive drugs, Diuretics with potassium, Thiazides and related diuretics.

AF: Atrial Fibrillation; BMI: body mass index; $CHADS_2$ and CHA_2DS_2-VASc: clinical prediction rules for estimating the risk of stroke based on Congestive heart failure, Hypertension, Age (≥65 = 1 point, ≥75 = 2 points), Diabetes, Stroke/TIA (2 points), and Vascular disease; HES: Hospital Episode Statistics; IQR: Interquartile range; ONS: Office for National Statistics; SD: standard deviation.

Chapter 3.3

Quality of INR control and outcomes following venous thromboembolism

Arlene M Gallagher, Frank de Vries, Jonathan M. Plumb, Bastian Haß, Andreas Clemens, Tjeerd-Pieter van Staa

Clin Appl Thromb Hemost. 2012 Jul;18(4):370-8

Abstract

Introduction: The objective of this study was to evaluate the pattern of anticoagulation after venous thromboembolism (VTE) in actual clinical practice.

Material and Methods: This study used the General Practice Research Database. Individuals aged 18+ years with VTE were matched to 3 controls.

Results: Of the 46,335 VTE cases and 138,024 controls, 70.2% of cases and 86.6% of controls had no obvious risk factors. The mortality risk was increased substantially around the time of diagnosis (relative hazard rate [RR] around 21) but remained elevated for further 4 years (RRs around 1.5-2.0). The mean percentage of time spent within the therapeutic range for International normalised ratio (INR) was 57.0%. The lowest rate of VTE recurrence occurred in patients with ≥70% time spent within therapeutic range (RR of 0.50, 95% CI 0.39-0.63 compared to <30%).

Conclusions: Higher time spent within therapeutic INR range was associated with lower risks of VTE recurrence and death due to VTE.

Keywords: coumarins, International Normalised Ratio, thromboembolism, VTE

Introduction

Venous thromboembolism (VTE), comprising deep venous thrombosis (DVT) and pulmonary embolism (PE), is a major cause of morbidity and mortality [1]. Venous thromboembolism can be classified into VTE caused by transient risk factors (such as surgery), permanent risk factors (such as cancer or paralysis) or without any identifiable risk factors (idiopathic VTE) [2].

Guidelines recommend acute treatment of VTE with low-molecular-weight heparin (LMWH) and dose-adjusted warfarin started immediately upon VTE diagnosis. Low-molecular-weight heparin should be discontinued once a therapeutic international normalised ratio (INR) has been achieved, usually defined as two back-to-back INR tests within therapeutic range (i.e. 2.0 to 3.0). Warfarin treatment should be continued for at least three months after which the physician should decide whether long-term secondary prevention should continue based on the risk of recurrence, risk of bleeding, and the preference of the patient to continue with warfarin treatment [3].

When first starting warfarin great effort is exerted to find the correct individual dose and achieve a stable therapeutic INR. As the recommended treatment duration is 3 months, this initiation phase may represent a substantial portion of the total treatment period. International normalised ratio levels outside the therapeutic range are associated with increased risk of VTE recurrence and bleeding. The objectives of this study were to evaluate the pattern of anticoagulation and quality of INR control after VTE in actual clinical practice and to assess the effects these have on VTE recurrence and mortality.

Methods

Source of data

Information for the study was obtained from the General Practice Research Database (GPRD), which comprises the computerised medical records of General Practitioners (GPs) in the UK. GPs play a key role in the UK health care system, as they are responsible for primary health care and specialist referrals. Patients are semi-permanently affiliated to a practice, which centralises the medical information from the GPs, specialist referrals, and hospitalisations. The data recorded in the GPRD include demographic information, prescription details, clinical events, preventive care provided, specialist referrals, hospital admissions and their major outcomes. Specialists and hospitals are required in the UK to inform the GP about the patient's medical care, and summary information from specialist care is typically entered into GPRD by the GP. The General Practice Research Database patients in English practices are now linked individually and anonymously to the national registry of hospital admission (Hospital Episode Statistics [HES]) and to the death certificates (as collected by the Office of National Statistics). For each hospitalised patient, the hospital charts are reviewed. The dates of admission and discharge, and main diagnoses are extracted, coded by coding staff and collated nationally. The death certificates list the date and causes of death. Hospital Episode Statistics data were available from 2001 to 2007 for 200 practices and death certificate data from 2001 to 2008. A high level of validity of VTE recording in GPRD has been reported previously [4, 5].

Study population

The study population consisted of patients aged 18 years or older with a diagnosis of VTE on or after January 01, 1995 and during the period of GPRD or HES data collection (until October 30, 2009). The index date was defined as the date of the first record of VTE from either source. Venous thromboembolism was defined as a composite of distal and proximal DVT, and PE. Patients with less than one year of GPRD data previous to their first VTE event were excluded. The risk of VTE is increased antepartum and 6 weeks postpartum. These events will not be managed with warfarin and will therefore be underrepresented. Patients with a record of pregnancy within 44 weeks before, or those with a delivery record in the 6 weeks after were excluded (2,912 patients). Information from the free text area was used to exclude patients negating or questionable free text (e.g. "?" or "ruled out"). Each patient with a VTE event was matched to up to three controls, by general practice, gender and year of birth (within five years). The pool of controls consisted of patients who had no record of VTE in GPRD or HES. The study population was followed from the index date until the end of data collection, patient's death or transfer out of the GP practice, whichever date came first.

In order to describe the aetiology of VTE, patients were classified into the following categories based on data recorded prior to the index date:

- Patients with a modifiable (transient) strong risk factor for VTE (category A). Hip or leg fracture, hip or knee replacement, major surgery (pelvis, lower leg, knee, feet, veins, arteries, spinal cord, unspecified region), major trauma (head, neck, spinal cord, trunk, pelvis, lower leg, knee, feet, unspecified region).
- Patients with a modifiable (transient) moderate / low risk factor for VTE (category B). Major surgery (head, neck, trunk, arms/hands), major trauma (arms/hands), pneumonia, chronic obstructive pulmonary disease, hormone therapy or oral contraceptives.
- Patients with an 'unmodifiable' risk factor for VTE (category C). Active cancer, congestive heart failure, varicose veins.
- Patients with no obvious risk factor (category D). All the other patients, not falling in the previous categories.

Conditional logistic regression was used to compare the medical history of VTE cases and controls, estimating the odds ratios (OR) and 95% confidence intervals (95% CI).

International Normalised Ratio and Percentage of Time Spent Within Therapeutic Range

Anticoagulation treatment requires frequent blood tests to monitor the level of anticoagulation. This is measured by the INR. International normalised ratio values are typically found in GPRD if the laboratory sends the results electronically to the general practice. The association between the rate of mortality and VTE recurrence and rate of discontinuation associated with the level of anticoagulation was evaluated during the time, with INR measurements evaluating the time spent within therapeutic range. This was done for VTE cases with at least three records of INR measurements. For each of these cases, the percentage of time spent within the therapeutic range (2.0 - 3.0) was calculated using the Rosendaal method of linear interpolation [6]. Patients were censored in this analysis if no repeat INR was measured within three months of a previous INR measurement, or at the patient's date of death, or end of data collection.

Treatment Persistence

Treated cases were identified as those with a prescription for coumarins (warfarin, acenocoumarol or phenindione) or LMWH or at least 2 INR measurements. For patients identified as being hospitalised, the index date was modified to that of the discharge date. For each coumarin prescription, the expected duration of use was taken as 28 days. Episodes of use were calculated by looking for repeat prescriptions in the period of time of expected duration of use plus 30 days.

Treatment persistence (i.e. repeat prescribing over time) immediately following VTE diagnosis was examined in those with a first-ever prescription for a coumarin on or within 30 days after the index date. Discontinuation was defined as no repeat prescription within 90 days after the expected end of the treatment course. Cox proportional hazards regression was used to estimate discontinuation with anticoagulant therapy over time. Fully adjusted models included (where appropriate) age, gender, BMI, smoking status, alcohol use, socioeconomic status at the location of the practice, hip, knee or lower leg fracture, hip or knee replacement, major trauma, recent surgery, active cancer, spinal cord injury, heart failure, chronic obstructive pulmonary disease, dementia, falls, oral contraceptive use, use of hormone therapy, or varicose veins.

Outcomes after VTE

The VTE cases were followed for the following three outcomes: (1) hospital re-admission for VTE after the initial VTE diagnosis (based on HES data), (2) death due to VTE (based on death certificate data) or (3) all-cause death. The study population was followed for these outcomes from the index date to the end of data collection, patient's death or transfer out of the practice, whichever date came first. Readmission for VTE was counted if a VTE was recorded at least 10 days after the index VTE event, as either the primary or secondary reason for hospitalisation. Death due to VTE was counted if VTE was recorded as one of the causes of death. Cox proportional hazards regression was used to estimate the relative hazard rates (RRs) of outcomes over time. Fully adjusted models included the same variables as listed in the section on treatment persistence. We also evaluated the patterns of the hazard rates (i.e., absolute risk) for death (based on the death certificates). The hazard rates were estimated by dividing the follow-up time into 100 periods (over five years) and by calculating the absolute rate within each small period (the hazard rate provides the risk of the outcome over a small period of time). These estimates were then smoothed using the methods proposed by Ramlau-Hansen [7].

Results

The study population included 46,335 VTE cases and 138,024 controls. Table 1 shows the baseline characteristics of the VTE cases and their matched controls. The mean age of VTE cases was 65 years and 55.2% of patients were female. Based on data recorded in GPRD, 70.2% of the VTE cases and 86.6% of controls had no obvious risk factors for VTE. The VTE cases were more likely to have a medical history of strong risk VTE factors compared to controls (OR 18.76; 95% CI 17.37 - 20.27). The occurrence of a recent fracture (6 weeks prior) of the lower extremity, hip or knee replacement or major trauma were all strongly associated with an increased risk of VTE (ORs of 32.92, 40.26 and 20.56, respectively). Recent use of hormone therapy or oral contraceptives was associated with VTE (OR of 1.42; 95% CI 1.34-1.49). Obese patients were also more likely to suffer VTE (OR of 2.01).

There were 20,090 VTE cases and 50,535 controls eligible for linkage to HES and 21,796 VTE cases and 64,673 controls eligible for linkage to the death certificate data. Table 2 shows the mortality rate and causes of death. The death certificates data showed that VTE cases were more likely to die compared to controls (RR of 5.00; 95 CI 4.76-5.24). Frequent causes of death in VTE cases were neoplasms (RR of 9.37) and diseases of the circulatory system (RR of 3.60). Figure 1 shows the crude RRs of mortality over time after diagnosis in VTE cases compared to controls. The hazard rate of death was increased substantially around the time of diagnosis (RRs around 21), decreased in the 12 months after VTE diagnosis and then remained elevated for a further four years (RRs around 1.5-2.0).

There were 5,162 VTE cases who completed the first three months of treatment. The large majority of VTE cases discontinued coumarin treatment in the next three months. Only 13% of the VTE cases (95% CI 12-14%) were found to have continued coumarin treatment for at least 9 months and 3% (95% CI 3-4%) for at least two years.

There were 10,381 VTE cases with at least three INR measurements recorded in GPRD (on average 23 measurements per patient (standard deviation 16). As shown in Table 3, the mean percentage of time spent within therapeutic range was 57.0%. VTE cases (aged 60 to 79 years) had the highest percentage spent within therapeutic range (59.2%) while youngest patients had the lowest (48.7% in patients aged <40 years). There was no substantive difference between women and men in the percentage of time spent within therapeutic range (58.0% and 56.0%, respectively).

Patients with VTE with a high percentage of time spent within therapeutic range were less likely to discontinue coumarin treatment compared to those with a low

percentage (Table 4). The adjusted RR in VTE cases with ≥70% time spent within therapeutic range was 0.53 (95% CI 0.48-0.59) compared to those with <30% time spent within therapeutic range.

Table 5 shows the outcomes after the VTE diagnosis. The VTE recurrence was found to be higher in elderly VTE cases compared to younger cases (RR of 1.53 for recurrent hospitalisation for VTE and RR of 11.01 for death due to VTE). The VTE recurrence was lower in VTE cases with strong modifiable (transient) risk factors compared to those with idiopathic VTE (RR of 0.84 for recurrent hospitalisation for VTE). The lowest rate of VTE recurrence occurred in patients with ≥70% time spent within therapeutic range (RR of 0.50, 95% CI 0.39-0.63 compared to those with <30% time spent within therapeutic range).

Figure 2 shows the life-table for VTE recurrence (based on recurrent hospitalisation for VTE). Readmission rates for VTE were high, especially in those patients with less time spent within therapeutic range. The incidence of VTE recurrence over two years was 11% in those with ≥70% and 21% in those with <30% time spent within therapeutic range. The largest risks for VTE recurrence occurred within the first three months after the first VTE.

Table 1: Baseline characteristics for VTE cases and controls

Characteristic	Cases N (%) (N=46,355)	Controls N (%) (N=138,024)	OR (95% CI)
Age			
18-39	4,132 (8.9%)	11,599 (8.4%)	-
40-59	11,202 (24.2%)	33,469 (24.2%)	-
60-79	21,289 (45.9%)	63,810 (46.2%)	-
80+	9,732 (21.0%)	29,146 (21.1%)	-
Gender			
Women	25,598 (55.2%)	75,795 (54.9%)	-
Men	20,757 (44.8%)	62,229 (45.1%)	-
BMI			
Underweight	2,180 (4.7%)	7,799 (5.7%)	1.02 (0.97 - 1.08)
Normal	11,046 (23.8%)	41,202 (29.9%)	Reference
Overweight	14,488 (31.3%)	43,690 (31.7%)	1.26 (1.22 - 1.29)[a]
Obese	12,175 (26.3%)	23,277 (16.9%)	2.01 (1.95 - 2.07)[a]
Unknown BMI	6,466 (13.9%)	22,056 (16.0%)	-
Smoking Status			
Non Smoker	20,981 (45.3%)	64,421 (46.7%)	Reference
Ex Smoker	12,041 (26.0%)	31,605 (22.9%)	1.21 (1.18 - 1.25)[a]
Smoker	8,984 (19.4%)	24,421 (17.7%)	1.13 (1.10 - 1.17)[a]
Unknown Smoking Status	4,349 (9.4%)	17,577 (12.7%)	-
Alcohol Drinking Status			
Non Drinker	6,394 (13.8%)	16,947 (12.3%)	Reference
Ex Drinker	2,677 (5.8%)	6,653 (4.8%)	1.09 (1.03 - 1.15)[a]
Drinker	28,889 (62.3%)	85,449 (61.9%)	0.89 (0.86 - 0.92)[a]
Unknown Drinking Status	8,395 (18.1%)	28,975 (21.0%)	-
Risk category for VTE prior to index			
Modifiable strong	4,298 (9.3%)	902 (0.7%)	18.76 (17.37 - 20.27)[a]
Modifiable moderate/low	6,449 (13.9%)	12,520 (9.1%)	2.03 (1.97 - 2.11)[a]
Unmodifiable	3,069 (6.6%)	5,043 (3.7%)	2.37 (2.26 - 2.49)[a]
No obvious risk factor	32,539 (70.2%)	119,559 (86.6%)	Reference

[a] p<0.01

BMI: body mass index; CI: confidence interval; OR: odds ratio; VTE: venous thromboembolism.

Table 2: Mortality rate and causes of death in VTE cases and controls

ICD-10 Chapter	Cause Description	VTE Cases (N=19,307)		Control Patients (N=57,464)		Crude RR (95% CI)
		N of Deaths	Incidence Rate#	N of Deaths	Incidence Rate#	
	All cause	4,530	169.8	2,648	34.3	5.00 (4.76 - 5.24) [a]
A00 - B99	Certain infectious and parasitic diseases	52	1.9	30	0.4	5.07 (3.24 - 7.95) [a]
C00 - D48	Neoplasms	2,278	85.4	712	9.2	9.37 (8.61 - 10.19) [a]
E00 - E90	Endocrine, nutritional and metabolic diseases	43	1.6	46	0.6	2.72 (1.79 - 4.12) [a]
F00 - F99	Mental and behavioural disorders	45	1.7	107	1.4	1.22 (0.86 - 1.73)
G00 - G99	Diseases of the nervous system	68	2.5	78	1.0	2.54 (1.83 - 3.51) [a]
I00 - I99	Diseases of the circulatory system[b]	1,227	46.0	996	12.9	3.60 (3.31 - 3.91) [a]
J00 - J99	Diseases of the respiratory system	414	15.5	379	4.9	3.18 (2.76 - 3.65) [a]
K00 - K93	Diseases of the digestive system	129	4.8	116	1.5	3.24 (2.52 - 4.16) [a]
L00 - L99	Diseases of the skin and subcutaneous tissue	17	0.6	6	0.1	8.27 (3.26 - 20.97) [a]
M00 - M99	Diseases of the musculoskeletal system and connective tissue	55	2.1	18	0.2	8.96 (5.26 - 15.25) [a]
N00 - N99	Diseases of the genitourinary system	59	2.2	55	0.7	3.12 (2.16 - 4.50) [a]
R00 - R99	Symptoms, signs and abnormal clinical and laboratory findings	38	1.4	51	0.7	2.16 (1.42 - 3.29) [a]
V01 - Y98	External causes of morbidity and mortality	92	3.4	48	0.6	5.63 (3.97 - 7.98) [a]

[a] $p<0.01$

[b] Includes MI, chronic ischaemic heart disease, pulmonary embolism, stroke, VTE.

per 1,000 person-years

CI: confidence interval; ICD-10: International Classification of Disease, version 10; MI: myocardial infarction; RR: Relative Rate; VTE: venous thromboembolism.

Table 3: Time spent within therapeutic range in venous thromboembolism (VTE) cases using coumarins

				Percentage of time spent within therapeutic range					
Category	Number of Patients	Mean	Median	<30 N (%)	30-39 N (%)	40-49 N (%)	50-59 N (%)	60-69 N (%)	≥70 N (%)
All	10,381	57.0	59.2	1,508 (14.5%)	975 (9.4%)	1,256 (12.1%)	1,594 (15.4%)	1,729 (16.7%)	3,319 (32.0%)
Age									
18-39	742	48.7	49.7	177 (23.9%)	89 (12.0%)	110 (14.8%)	112 (15.1%)	101 (13.6%)	153 (20.6%)
40-59	2,351	57.1	59.7	349 (14.8%)	212 (9.0%)	269 (11.4%)	357 (15.2%)	401 (17.1%)	763 (32.5%)
60-79	5,112	59.2	61.4	626 (12.2%)	438 (8.6%)	605 (11.8%)	774 (15.1%)	868 (17.0%)	1,801 (35.2%)
80+	2,176	54.5	56.4	356 (16.4%)	236 (10.8%)	272 (12.5%)	351 (16.1%)	359 (16.5%)	602 (27.7%)
Gender									
Women	4,900	58.0	60.3	674 (13.8%)	428 (8.7%)	588 (12.0%)	738 (15.1%)	818 (16.7%)	1,654 (33.8%)
Men	5,481	56.0	58.1	834 (15.2%)	547 (10.0%)	668 (12.2%)	856 (15.6%)	911 (16.6%)	1,665 (30.4%)

VTE: venous thromboembolism.

Table 4: Discontinuation rate with coumarins over the first two years after VTE diagnosis stratified by the percentage of time spent within therapeutic range (N = 4,337)

Percentage of time spent within therapeutic range	N (%)	N discontinue	Discontinuation rate (%) over 2 years (95% CI)	Age and gender adjusted RR (95% CI)	Fully adjusted RR (95% CI)
<30	571 (13.2%)	487	0.96 (0.94 - 0.98)		
30-39	395 (9.1%)	339	0.94 (0.91 - 0.96)	0.82 (0.71 - 0.94) [a]	0.82 (0.71 - 0.94) [a]
40-49	510 (11.8%)	432	0.92 (0.90 - 0.95)	0.66 (0.58 - 0.75) [a]	0.66 (0.58 - 0.76) [a]
50-59	687 (15.8%)	568	0.90 (0.87 - 0.92)	0.57 (0.51 - 0.65) [a]	0.57 (0.51 - 0.65) [a]
60-69	740 (17.1%)	625	0.89 (0.87 - 0.92)	0.56 (0.49 - 0.63) [a]	0.56 (0.49 - 0.63) [a]
≥70	1,434 (33.1%)	1174	0.87 (0.85 - 0.89)	0.53 (0.48 - 0.59) [a]	0.53 (0.48 - 0.59) [a]

CI: confidence interval; RR: relative hazard rate; VTE: venous thromboembolism.

[a] $p<0.01$

Table 5: Outcomes after VTE (recurrence and death) stratified by age, gender, presence of VTE risk factors and percentage of time spent within therapeutic range

	VTE Recurrence with Hospitalisation (Based on HES Records)		
Characteristic	**N Recurrent Cases**	**Incidence Rate per 1,000 person-years**	**Adjusted[c] RR (95% CI)**
All	2,978	97	-
Age			
18-39	234	84.9	Reference
40-59	602	75.2	1.07 (0.74 - 1.55)
60-79	1,384	95.4	1.26 (0.89 - 1.79)
80+	758	139.7	1.53 (1.06 - 2.21) [a]
Gender			
Women	1,326	97.9	Reference
Men	1,652	96.3	1.03 (0.88 - 1.21)
Risk factor for VTE			
Modifiable strong	332	75.6	0.84 (0.66 - 1.07)
Modifiable moderate/low	502	115.1	1.15 (0.93 - 1.43)
Unmodifiable	226	125.3	1.46 (1.10 - 1.94) [b]
No obvious risk factor	1,918	95.2	Reference
Percentage Time in Range			
<30%	136	160.6	Reference
30-39%	77	124.9	0.80 (0.60 - 1.06)
40-49%	74	93.1	0.63 (0.47 - 0.83) [b]
50-59%	93	88.2	0.61 (0.47 - 0.79) [b]
60-69%	118	103.5	0.72 (0.56 - 0.92) [b]
≥70%	161	68.9	0.50 (0.39 - 0.63) [b]

[a] $p<0.05$

[b] $p<0.01$

[c] Adjusted for age, gender, BMI, smoking, alcohol, VTE risk category.

BMI: body mass index; CI: confidence interval; HES: Hospital Episode Statistics; RR: relative hazard rate; VTE: venous thromboembolism.

Table 5 (cont.): Outcomes after VTE (recurrence and death) stratified by age, gender, presence of VTE risk factors and percentage of time spent within therapeutic range

	Death Due to VTE (Based on Death Certificates)		
Characteristic	**N deaths**	**Incidence Rate per 1,000 person-years**	**Adjusted[c] RR (95% CI)**
All	997	37.4	-
Age			
18-39	17	6.5	Reference
40-59	91	13.2	2.17 (1.29 - 3.65) [b]
60-79	406	33.2	5.11 (3.13 - 8.33) [b]
80+	483	98.4	11.01 (6.75 - 17.96) [b]
Gender			
Women	420	34.9	Reference
Men	577	39.4	0.88 (0.77 - 1.01)
Risk factor for VTE			
Modifiable strong	116	31.7	1.03 (0.84 - 1.26)
Modifiable moderate/low	197	53.4	1.62 (1.38 - 1.92) [b]
Unmodifiable	132	89.2	1.72 (1.42 - 2.08) [b]
No obvious risk factor	552	30.9	Reference
Percentage			
Time in Range			
<30%	13	15	Reference
30-39%	8	12.4	0.76 (0.31 - 1.85)
40-49%	4	5	0.31 (0.10 - 0.95) [a]
50-59%	10	9.3	0.57 (0.25 - 1.30)
60-69%	6	5.2	0.31 (0.12 - 0.82) [a]
≥70%	16	7.1	0.44 (0.21 - 0.92) [a]

[a] p<0.05

[b] p<0.01

[c] Adjusted for age, gender, BMI, smoking, alcohol, VTE risk category.

BMI: body mass index; CI: confidence interval; HES: Hospital Episode Statistics; RR: relative hazard rate; VTE: venous thromboembolism.

Table 5 (cont.): Outcomes after VTE (recurrence and death) stratified by age, gender, presence of VTE risk factors and percentage of time spent within therapeutic range

	Death (Based on Death Certificates)		
Characteristic	**N VTE deaths**	**Incidence Rate per 1,000 person-years**	**Adjusted[c] RR (95% CI)**
All	4,506	169.3	-
Age			
18-39	76	29.1	Reference
40-59	543	78.7	2.93 (2.30 - 3.72) [b]
60-79	2,198	179.9	6.48 (5.14 - 8.15) [b]
80+	1,689	345.7	10.06 (7.98 - 12.70) [b]
Gender			
Women	2,031	169.2	Reference
Men	2,475	169.4	0.88 (0.83 - 0.94) [b]
Risk factor for VTE			
Modifiable strong	521	142.6	0.94 (0.85 - 1.03)
Modifiable moderate/low	909	246.6	1.54 (1.42 - 1.66) [b]
Unmodifiable	477	323.8	1.54 (1.39 - 1.70) [b]
No obvious risk factor	2,599	146.0	Reference
Percentage			
Time in Range			
<30%	168	194	Reference
30-39%	104	161.3	0.80 (0.63 - 1.03)
40-49%	99	124.6	0.63 (0.49 - 0.81) [b]
50-59%	104	96.5	0.49 (0.39 - 0.63) [b]
60-69%	87	75.8	0.37 (0.29 - 0.48) [b]
≥70%	122	54.2	0.28 (0.22 - 0.36) [b]

[a] p<0.05

[b] p<0.01

[c] Adjusted for age, gender, BMI, smoking, alcohol, VTE risk category.

BMI: body mass index; CI: confidence interval; HES: Hospital Episode Statistics; RR: relative hazard rate; VTE: venous thromboembolism.

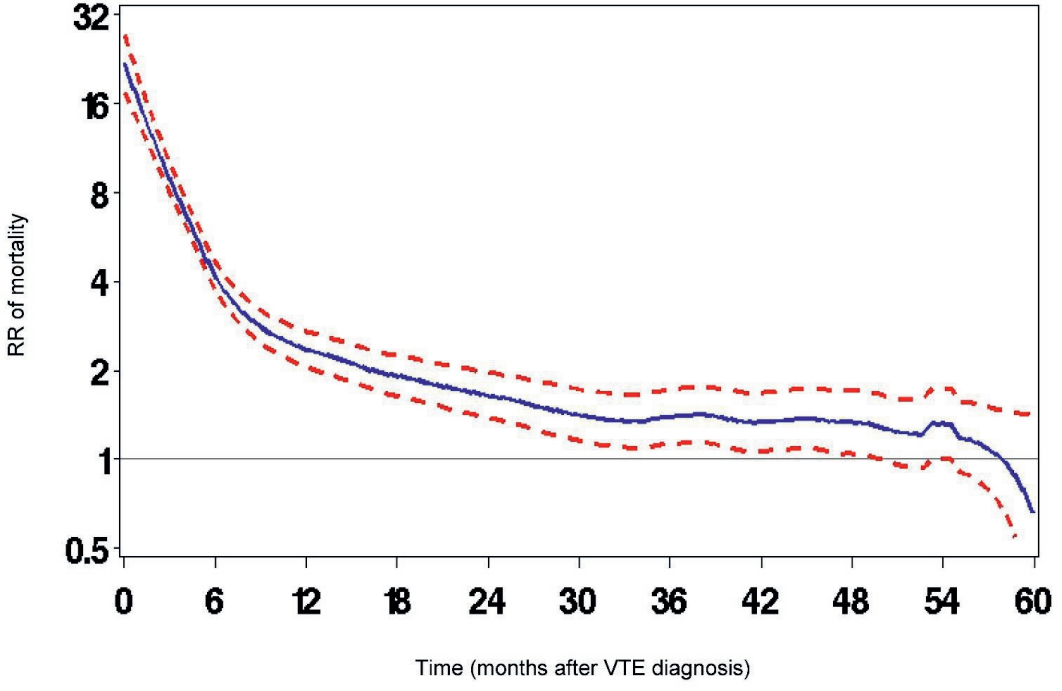


Figure 1: Crude RRs of mortality over time after diagnosis in VTE cases compared to controls (dotted lines indicate the 95% CI)

X-axis: Time (months after VTE diagnosis)
Y-axis: RR of mortality in VTE cases compared to controls
RR: relative hazard rate; VTE: venous thromboembolism.

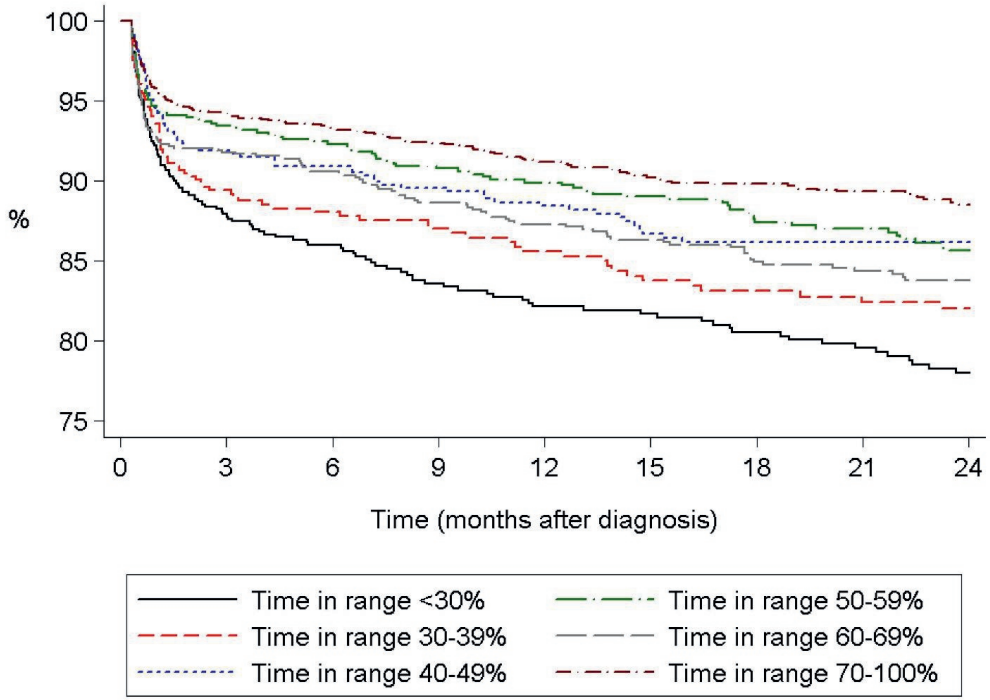


Figure 2: Percentage of VTE cases without recurrence of VTE (based on hospitalisation data in HES) stratified by percentage of time spent within therapeutic range

X-axis: Time (months after VTE diagnosis); Y-axis: Percentage of VTE cases without recurrence of VTE
HES: Hospital Episode Statistics; VTE: venous thromboembolism.

Appendix A

ICD-10 codes used to define Venous Thromboembolism (VTE)

ICD-10 code	ICD-10 term
I26.0	Pulmonary embolism with mention of acute cor pulmonale
I26.9	Pulmonary embolism without mention of acute cor pulmonale
I80.1	Phlebitis and thrombophlebitis of femoral vein
I80.2	Phlebitis/thrombophlebitis oth deep vessels low extremties
I82.2	Embolism and thrombosis of vena cava
I82.3	Embolism and thrombosis of renal vein
I82.8	Embolism and thrombosis of other specified veins
I82.9	Embolism and thrombosis of unspecified vein

ICD-10: International Classification of Disease, version 10.

Discussion

This study found that the risk of VTE was substantially increased in patients with a recent lower extremity fracture, hip or knee replacement or major trauma. The risk of mortality was substantially higher in VTE cases compared to controls. Recurrence of VTE (as measured by readmission to the hospital) was high and inversely related to the time spent within therapeutic range. Most patients with VTE stopped coumarin treatment within six months, in line with treatment guidelines, and a clear relationship exists between treatment discontinuation and time spent within therapeutic range.

Various studies in the literature have evaluated risk factors for the development and recurrence of VTE. The results in this study on risk factors are broadly consistent with the body of evidence in the literature. This study confirmed the strong associations between the development of VTE and various transient risk factors (such as major trauma and surgery) [2]. Obese patients in GPRD also had an increased risk of developing VTE similar to that previously reported [8]. However, some of our results on VTE recurrence differed with those of various published studies. Several studies have reported considerably higher risks of VTE recurrence in patients with cancer compared to those with transient risk factors [9, 10]. While we did find a higher risk in cancer patients, the excess risk was small. This study also found no major difference in VTE recurrence between men and women. A recent meta-analysis found that men had a 50% higher risk of VTE recurrence compared to women [11]. There may be several reasons for these discrepancies (other than bias). One reason may be differences in characteristics and representativeness of the study populations. Many of the studies reported in literature were conducted in specialist centres, while this study was conducted in routine clinical practice.

High recurrence rates of VTE have been reported in other observational studies conducted in non-specialist settings [9, 12, 13]. The largest study to date, including 2,090 VTE cases as recorded in a US claims database, found a recurrence rate of 10.9% after 6 months [12]. However lower rates of VTE recurrence have been reported in studies conducted in specialist anticoagulation centres [14, 15]. Lower rates of VTE rates have also been reported in randomised clinical studies. A recent large clinical study comparing dabigatran and warfarin found a VTE recurrence rate of about 2% over six months [16]. There has been limited research on the possible reasons for these differences in VTE recurrence rates and whether these differences may be explained by varying patterns of anticoagulation. There may be multiple and heterogeneous reasons for suboptimal anticoagulation with warfarin. These include, among others, co-prescribing of drugs that can interact with warfarin. The list of drugs with a potential to interact with warfarin is continuously increasing. Furthermore, complex dosage instructions may adversely affect patient compliance.

Hopefully, the next generation of oral anticoagulants that do not require anticoagulation monitoring may help to improve patient compliance.

Patients treated with coumarins are regularly monitored in order to maintain INR values within the range of 2.0 to 3.0. A recent US study that reviewed medical charts of VTE cases found that 37.7% of time was spent within therapeutic range [12]. We found that VTE cases treated in the UK with coumarins and with at least three INR measurements spent on average 57.0% of time within the therapeutic range. Direct comparisons between studies may be difficult given the differences in data collection and methods. Furthermore, the present study found that time spent within therapeutic range was a strong predictor of the discontinuation rate with coumarins. This could mean a ‘‘positive survivor bias’’ and an artificial high value in our study. Unfortunately, duration of treatment with coumarins was not reported in the US study. Time spent within therapeutic range was strongly associated with risk of death due to VTE and VTE recurrence. Lower risks were found in those with ≥70% time spent within therapeutic range. The strong association between level of anticoagulation with coumarins and adverse outcomes (stroke and mortality) has also been found in patients with atrial fibrillation [17]. A recent study found that patients have substantial difficulties maintaining adequate adherence with coumarins, and that this poor adherence has a significant effect on anticoagulation control [18]. In the present study, we did not have information on the reasons for poor adherence in some of treated VTE cases.

The strengths of this study were that the study population was representative of actual clinical practice. This study is the first VTE study using the GPRD with episodic hospital data linked to longitudinal data from general practice. However, there are also several important limitations of this study. We did not have information on the diagnostic criteria used for the VTE diagnosis, however GPRD has been used in previous studies evaluating VTE and a high level of validity in the recording of VTE by GPs has been reported [4, 5]. Also, we did not have the reasons for VTE recurrence. We may have over-estimated the rate of recurrence by assuming that further VTE records after 10 days constituted a new event. It is possible that some of these events were pre-existing diagnoses being re-recorded, which might inflate recurrence rates. Another limitation is that we did not have information on all risk factors for VTE, such as frailty. Also, like in any epidemiological study, the comparison groups in this study were not randomised. It could not be established whether unstable INR measurements caused the increased risk of mortality and VTE recurrence or whether instability in INR measurements was a symptom of intercurrent illness. We did not have information on INR values for all patients as the monitoring may have been performed elsewhere and not all anticoagulation clinics report INR results electronically to the GPs. It may be that results falling outside of therapeutic range are more likely to be transmitted in this way. Moreover, applying

the Rosendaal method to patients with only 3 INR measures has the potential to give a biased picture of INR control compared to patients with a larger number of test results. For example, a patient with three consecutive, out of range INR tests, taken on the first day of the month over three months would have a percentage time within therapeutic range of zero, whereas if the second test was in range the percentage would be 50%. Potentially, distal DVT can be managed in the outpatient setting which could have led to an underestimation of the recurrent VTE rate since our analysis only included recurrent events that required hospitalisation. Another limitation concerned the use of death certificates for classification for cause of death. One study found discrepancies between causes of death as reported on death certificates and those found at necropsy [19].

In conclusion, most VTE cases studied were treated with coumarins for less than six months and the overall time spent within therapeutic range was relatively low. Higher time spent within therapeutic range was associated with lower risks of VTE recurrence, lower discontinuation rates, and death due to VTE. The challenge remains to improve the time spent within the recommend INR range, but this may be demanding since the time required to achieve a stable therapeutic INR represents a substantial portion of the typical duration of treatment.

References

1. Kyle PA, Eichringer S. Deep venous thrombosis. Lancet. 2005;365:1163-74.
2. Samama MM. An epidemiologic study of risk factors for deep vein thrombosis in medical outpatients: the Sirius study. Arch Intern Med. 2000;160:3415-20.
3. Büller HR, Agnelli G, Hull RD, *et al*. Antithrombotic therapy for venous thromboembolic disease: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest. 2004;126:401S-428S.
4. Lawrenson R, Todd JC, Leydon GM, Williams TJ, Farmer RD. Validation of the diagnosis of venous thromboembolism in general practice database studies. Br J Clin Pharmacol. 2000;49:591-6.
5. Huerta C, Johansson S, Wallander MA, García Rodríguez LA. Risk factors and short-term mortality of venous thromboembolism diagnosed in the primary care setting in the United Kingdom. Arch Intern Med. 2007;167:935-43.
6. Rosendaal FR, Cannegieter SC, van der Meer FJ, Briët E. A method to determine the optimal intensity of oral anticoagulant therapy. Thromb Haemost. 1993;69:236-9.
7. Ramlau-Hansen H. Smoothing counting process intensities by means of kernel functional. Ann Statist. 1983;11:453–66.
8. Eichinger S, Hron G, Bialonczyk C, *et al*. Overweight, obesity, and the risk of recurrent venous thromboembolism. Arch Intern Med. 2008;168:1678-83.
9. Heit JA, Mohr DN, Silverstein MD, *et al*. Predictors of recurrence after deep vein thrombosis and pulmonary embolism: a population-based cohort study. Arch Intern Med. 2000;160:761-8.
10. Cushman M, Tsai AW, White RH, *et al*. Deep vein thrombosis and pulmonary embolism in two cohorts: the Longitudinal Investigation of Thromboembolism Etiology. Am J Med. 2004;117:19–25.
11. McRae S, Tran H, Schulman S, *et al*. Effect of patient's sex on risk of recurrent venous thromboembolism: a meta-analysis. Lancet. 2006;368:371-8.
12. Willey VJ, Bullano MF, Hauch O, *et al*. Management patterns and outcomes of patients with venous thromboembolism in the usual community practice setting. Clin Ther. 2004;26:1149-59.
13. Spencer FA, Gore JM, Lessard D, *et al*. Patient outcomes after deep vein thrombosis and pulmonary embolism: the Worcester Venous Thromboembolism Study. Arch Intern Med. 2008;168:425-30.
14. Harrison L, McGinnis J, Crowther M, *et al*. Assessment of outpatient treatment of deep-vein thrombosis with low-molecular-weight heparin. Arch Intern Med. 1998;158:2001-3.
15. Tillman DJ, Charland SL, Witt DM. Effectiveness and economic impact associated with a program for outpatient management of acute deep vein thrombosis in a group model health maintenance organization. Arch Intern Med. 2000;160:2926-32.
16. Schulman S, Kearon C, Kakkar AK, *et al*; RE-COVER Study Group. Dabigatran versus warfarin in the treatment of acute venous thromboembolism. N Engl J Med. 2009;361:2342-52.
17. Morgan CL, McEwan P, Tukiendorf A, *et al*. Warfarin treatment in patients with atrial fibrillation: Observing outcomes associated with varying levels of INR control. Thromb Res. 2009;124:37-41.
18. Kimmel SE, Chen Z, Price M, *et al*. The influence of patient adherence on anticoagulation control with warfarin: results from the International Normalized Ratio Adherence and Genetics (IN-RANGE) Study. Arch Intern Med. 2007;167:229-35.
19. Sington JD, Cottrell BJ. Analysis of the sensitivity of death certificates in 440 hospital deaths: a comparison with necropsy findings. J Clin Pathol 2002;55(7):499-502.

Chapter 3.4

Risk of death and cardiovascular outcomes with thiazolidinediones: a study with the General Practice Research Database and secondary care data

Arlene M Gallagher, Liam Smeeth, Suzie Seabroke,
Hubert GM Leufkens, Tjeerd P van Staa

PLoS One. 2011;6(12):e28157
Erratum in: PLoS One. 2013;8(2)

Abstract

Objective: To describe the likely extent of confounding in evaluating the risks of cardiovascular (CV) events and mortality in patients using diabetes medication.

Methods: The General Practice Research Database was used to identify inception cohorts of insulin and different oral antidiabetics. An analysis of bias and incidence of mortality, acute coronary syndrome, stroke and heart failure were analysed in GPRD, Hospital Episode Statistics and death certificates.

Results: 206,940 patients were identified. The bias analysis showed that past thiazolidinedione users had a lower mortality risk compared to past metformin users. There were no differences between past users of rosiglitazone and pioglitazone (adjusted RR of 1.04; 95% CI 0.93-1.18). Current rosiglitazone users had an increased risk of death (adjusted RR 1.20; 95% CI 1.08-1.34) and of hospitalisation for heart failure (adjusted RR of 1.73; 95% CI 1.19-2.51) compared to current pioglitazone users. Risk of mortality was increased two-fold shortly after starting rosiglitazone. Excess risk of death over 3 years with rosiglitazone was 0.3 per 100 in those aged 50-64 years, 2.0 aged 65-74, 3.0 aged 75-84, and 7.0 aged 85+. The cause of death with rosiglitazone was more likely to be due to a disease of the circulatory system.

Conclusions: Higher risks for death (overall and due to cardiovascular disease) and heart failure were found for rosiglitazone compared to pioglitazone. These excess risks were largest in patients aged 65 years or older. The European regulatory decision to suspend rosiglitazone is supported by this study.

Introduction

The thiazolidinediones - rosiglitazone and pioglitazone - have been widely used for the treatment of type 2 diabetes mellitus. However, signals concerning the cardiovascular safety have been emerging over the last few years. An increased risk of myocardial infarction (MI), with no increases in mortality was found for rosiglitazone in a meta-analysis of randomised controlled trials (RCTs) [1, 2]. Pioglitazone was reported to have a statistically significant lower risk in a composite endpoint of death, MI and stroke from findings of another meta-analysis [3]. These reports were supported by a FDA meta-analysis of RCTs, which found that the risk of all-cause and cardiovascular mortality and MI tended to be lower with pioglitazone and higher with rosiglitazone (although most results did not reach statistical significance) [4]. This analysis also highlighted that the risk of congestive heart failure was increased with both drugs [4]. Another study also noted the increase in numbers of congestive heart failure with both thiazolidinediones [5]. Limitations of these meta-analyses are the low number of events and the pooling of studies with varying designs and study populations. Furthermore, the populations using the drugs and exposure characteristics in actual clinical practice may be different from those enrolled in RCTs [6].

A large number of observational studies have also evaluated the cardiovascular safety of rosiglitazone or pioglitazone but the quality of these studies varied considerably [7-32]. An unpublished assessment by the UK medicines and medical devices regulatory authority of 24 observational studies noted substantial limitations in most studies, including limited statistical power, short durations of follow-up or study populations with lack of long-term continuity of data collection. Also, observational studies relied on the power of statistical adjustment to deal with confounding. Confounding (which is a challenge in any observational study) may be even more challenging in studies of diabetes, since drug exposure is defined by diabetes severity, making it very difficult to separate the effect of disease severity from treatment.

There has been an ongoing European regulatory review of the cardiovascular safety of thiazolidinediones and a study was commissioned by the UK Medicines and Healthcare products Regulatory Agency (MHRA) (this was done prior to the subsequent suspension of rosiglitazone in Europe). The first objective of the present study was to describe the likely extent of confounding in evaluating the risks of cardiovascular events and mortality in patients using diabetes medication. Where confounding effects could be adequately controlled, the second objective was to compare the risks of cardiovascular events and mortality between different types of diabetes medication, including rosiglitazone and pioglitazone.

Methods

Data source

This study used data from the General Practice Research Database (GPRD) in the United Kingdom. GPRD comprises the computerised medical records maintained by general practitioners (GPs). GPs play a key role in the UK health care system, as they are responsible for primary health care and specialist referrals. Patients are affiliated with a practice, which centralises the medical information from the GPs, specialist referrals and hospitalisations. The data recorded in the GPRD since 1987 include demographic information, prescription details, clinical events, preventive care provided, specialist referrals, hospital admissions and their major outcomes [33]. A recent review of all validation studies found that medical data in GPRD were generally of high quality [34]. Patients in about 40% of GPRD have now been linked individually and anonymously to the national registry of hospital admission (Hospital Episode Statistics [HES]) and to the death certificates (as collected by the Office of National Statistics). For each hospitalised patient, the hospital charts are reviewed, dates of admission and discharge and main diagnoses are extracted, coded by coding staff and collated nationally into HES. The death certificates list the date and causes of death. HES data were available from April 1997 and death certificates from January 2001 for about 40% of GPRD practices. The data from HES and GPRD were recorded and collected independently from each other.

Study population

The exposed study cohort consisted of adults aged 40 years and older with a prescription for insulin or oral antidiabetic drugs (OAD) at least one year after start of data collection. Patients with a record of type I diabetes were excluded. The index date was the first prescription for insulin or OAD one year after start of GPRD data collection (the prescribing prior to the index date was not considered in the creation of this prevalent user cohort). The period of follow-up was from the index date up to the date of censoring (i.e., transfer out of the practice, last collection from the practice, or death). Each exposed patient was matched by age (within 5 years), sex and practice to one control patient, with the index date of the control being the same as that of the exposed patient. Within this overall exposed cohort, we identified inception cohorts for each class of diabetes medication (a patient was included in an inception cohort if they received first-ever prescription for a class of diabetes medication at least one year after start of GPRD data collection). The medications of interest in this study were thiazolidinediones, insulins, metformin and sulphonylureas. Furthermore, we also created two inception cohorts separately for pioglitazone and rosiglitazone. Patients prescribed multi-constituent preparations were included in multiple classes of diabetes medication.

In order to prevent immortal time bias and incorrect exclusion of patients based on events that occur after the index date, patients could belong to multiple inception cohorts. In comparisons between different diabetes medications, patients were censored at the start of treatment with the medication of the reference group. The exposure to each diabetes medication was classified in a time-dependent manner, dividing the period of follow-up into periods of current, recent and past exposure. The period of current exposure was defined as the period from the date of a prescription up to 3 month after the date of the prescription. Recent use was the period of time from 3 to 12 months after the most recent prescription and past use was the time from 12 months after. Patients could move between exposure categories over time.

Outcomes of interest

The following incident outcomes were measured: death due to any cause (as recorded in GPRD or on death certificates), cause of death (death certificates), acute coronary syndrome (ACS) (recorded in GPRD or HES), stroke (GPRD or HES) and heart failure (GPRD or HES). Analyses requiring the HES or ONS data were restricted to patients from practices participating in the linkage and to those with data during the HES/ONS data collection period. Given the different coding dictionaries used by the various datasets and different methods for data collection, outcomes from each source were analysed separately.

Statistical analyses

Four sets of analyses were conducted. The first set of analyses evaluated the possible extent of bias in the comparisons between different diabetes medications and whether statistical adjustment with risk factors would sufficiently address any confounding. In this analysis, the incidence of various outcomes was compared between past exposure for each of the diabetes medications and matched control patients. Poisson regression was used to estimate relative rates (RRs). These models also included age, sex, calendar year, small-area socioeconomic status (for linked practices), smoking status, use of alcohol, body mass index, medical history ever before of coronary heart disease, coronary revascularisation, hyperlipidaemia, hypertension, peripheral vascular disease, renal impairment and stable angina and prescribing in the 6 months before of angiotensin II receptor blockers, antiplatelets, beta blockers, calcium channel blockers, diuretics, nitrates, NSAIDs or aspirin and statins. In addition, the models included current use of the various classes of diabetes medication. Missing values for alcohol use, smoking status and body mass index were included as separate categories in the regression analyses. This bias analysis explored whether statistical adjustment substantially reduced the point estimates of RRs towards one. If the adjusted RRs would remain elevated, this could either indicate residual confounding (i.e., effects of the underlying disease) or persistent adverse effects after treatment discontinuation.

The second set of analyses concerned a comparison of the rates of outcomes during current use of different diabetes medications. These analyses were restricted to the types of diabetes medications that did not have major differences in risk during past use (as observed in the previous set of analyses). These analyses were stratified by age, co-prescribing of insulin and calendar time (before and after 2007; in October 2007 additional warnings for cardiovascular disease with rosiglitazone were communicated by the UK regulatory authority.

The third set of analyses described the pattern of risks over duration of treatment. The follow-up period of current exposure was divided into 100 periods and the absolute risk was estimated within each small period. These estimates were then smoothed using the method proposed by Ramlau-Hansen [35]. This analysis of hazard rates displays visually the observed (crude) risks over duration of current exposure to a diabetes medication.

The last set of analyses estimated the cumulative incidence over time with current use of various medications for diabetes. Kaplan-Meier life-tables were estimated. These life-tables describe the absolute incidence over time, accounting for loss to follow-up while not adjusting for any risk factors.

Results

Demographics

The overall study population included 206,940 patients prescribed insulin or OAD and the same number of controls without diabetes. Table 1 shows the baseline characteristics of the inception cohorts of metformin, sulphonylureas, rosiglitazone, pioglitazone and insulins. As expected, there were differences in risk factors between metformin, sulphonylureas and insulin (metformin was more often the first diabetes treatment while insulin users had more frequent history of use of other diabetes treatments). There were no major differences at baseline for most characteristics between rosiglitazone and pioglitazone, except for prescribing over calendar time and higher prior use of statins among pioglitazone users (this was found to be related to secular changes in statin prescribing). The number of patients starting rosiglitazone dropped substantially after 2007 (2006, N=4583; 2007, N=2612; 2008, N=474; 2009, N=278). The use of pioglitazone changed differently (2006, N=1543; 2007, N=2815; 2008, N=4660; 2009, N=3177).

Bias analyses

Table 2 shows the results of the bias analyses. The statistical comparison showed major differences, as expected, in the risks of cardiovascular outcomes and death between the overall exposed cohort and control cohort. Statistical adjustment did reduce the RRs for the cardiovascular outcomes, although statistically significant differences remained. For mortality, statistical adjustment increased the RRs with higher risks in the exposed cohort (indicating that diabetes treatments are less likely to be given to patients at imminent risk of death). In the analysis of past exposure, it was found that patients who had discontinued insulin had a higher risk of death compared to past users of metformin. Past thiazolidinediones users had a lower risk of death. The RRs were comparable in past rosiglitazone users compared to past pioglitazone users (adjusted RRs of 1.04 [95% CI 0.93-1.18], 1.11 [95% CI 0.86-1.43], 0.76 [95% CI 0.56-1.03] and 0.95 [95% CI 0.74-1.23] for, respectively, death, ACS, stroke and heart failure in GPRD). These results suggest that comparisons of different classes of diabetes medications are likely to be prone to substantial confounding, while the within class comparison of rosiglitazone versus pioglitazone is less prone to selection bias and confounding.

Comparison of rates of outcomes with current diabetic medications

As shown in Table 3, current rosiglitazone users had an increased risk of death compared to current pioglitazone users (adjusted RR 1.20 [95% CI 1.08-1.34]). The rates of ACS and stroke were comparable between rosiglitazone and pioglitazone. Hospital admission for congestive heart failure was increased with rosiglitazone (adjusted RR of 1.73 [95% CI 1.19-2.51]) while heart failure diagnosed or treated by

a GP was statistically comparable between the two types of thiazolidinediones (adjusted RR of 1.14 [95% CI 0.97-1.34]). Table 4 shows the analyses stratified by age, co-prescribing of insulin and calendar time. The RRs comparing current rosiglitazone to pioglitazone users did not vary substantially.

Table 5 shows the mortality rates and primary cause of death in current rosiglitazone and pioglitazone users. Rosiglitazone users were more likely to die due to a disease of the circulatory system compared to pioglitazone users (although these findings did not reach statistical significance).

Patterns of Risks

Figure 1 shows the smoothed RRs over duration of treatment in rosiglitazone users compared to pioglitazone users. It was found that the risk of mortality was increased about two-fold shortly following the start of rosiglitazone treatment (compared to pioglitazone) and then remained elevated (although at a much lower level) with longer treatment duration.

Cumulative incidence over time

An additional seven patients (per 100 patients over 3 years) died during rosiglitazone treatment compared to pioglitazone in those aged 85 years or older (Table 6). The excess risks were progressively smaller in younger patients (there was an excess risk of death of 0.3 per 100 in those aged 65 years or younger). Table S1 (supporting information) lists the cumulative incidence of ACS, stroke and heart failure.

Table 1: Baseline characteristics at inception date of metformin, sulphonylureas, rosiglitazone, pioglitazone and insulin

Characteristic	Metformin N = 121,637 N (%)	Sulphonylureas N = 76,863 N (%)	Rosiglitazone N = 22,636 N (%)	Pioglitazone N = 18,953 N (%)	Insulin N = 26,458 N (%)
Mean Age, years (SD)	64 (12)	66 (12)	63 (11)	64 (11)	66 (11)
Median Age, years (IQR)	64 (55-73)	66 (57-74)	63 (55-72)	64 (56-72)	66 (57-74)
Sex:					
Female	53,649 (44.1%)	33,982 (44.2%)	9,939 (43.9%)	7,939 (41.9%)	11,678 (44.1%)
Male	67,988 (55.9%)	42,881 (55.8%)	12,697 (56.1%)	11,014 (58.1%)	14,780 (55.9%)
Mean Follow-up, years (SD)	4 (3)	5 (4)	5 (2)	3 (2)	4 (3)
Mean BMI (SD)	31 (6)	30 (6)	31 (6)	32 (6)	30 (6)
Mean HbA1c,% (SD)	8.8 (1.8)	9.0 (1.8)	8.9 (1.5)	8.7 (1.5)	9.8 (1.8)
Smoking Status					
Non Smoker	48,487 (39.9%)	30,748 (40.0%)	8,736 (38.6%)	6,799 (35.9%)	9,941 (37.6%)
Ex Smoker	46,146 (37.9%)	26,809 (34.9%)	9,556 (42.2%)	8,989 (47.4%)	10,421 (39.4%)
Smoker	22,330 (18.4%)	13,760 (17.9%)	3,950 (17.5%)	3,031 (16.0%)	4,909 (18.6%)
Unknown	4,674 (3.8%)	5,546 (7.2%)	394 (1.7%)	134 (0.7%)	1,187 (4.5%)
Hospitalisation in the year before	11,391 (22.6%)	8,658 (27.4%)	1,979 (20.5%)	1,747 (22.0%)	5,112 (43.8%)
Number of diabetes medication classes ever before:					
0	89,922 (73.9%)	36,260 (47.2%)	645 (2.8%)	389 (2.1%)	2,917 (11.0%)
1	29,067 (23.9%)	34,393 (44.7%)	8,628 (38.1%)	4,595 (24.2%)	3,081 (11.6%)
2	2,371 (1.9%)	5,703 (7.4%)	11,302 (49.9%)	8,747 (46.2%)	11,024 (41.7%)
3	248 (0.2%)	459 (0.6%)	1,838 (8.1%)	4,357 (23.0%)	7,683 (29.0%)
4	29 (0.0%)	45 (0.1%)	207 (0.9%)	746 (3.9%)	1,563 (5.9%)
5+	0 (0.0%)	3 (0.0%)	16 (0.1%)	119 (0.6%)	190 (0.7%)

ACE: angiotensin-converting-enzyme inhibitors; ACS: Acute Coronary Syndrome; BMI: Body Mass Index; HbA1c: Haemoglobin A1C test; NSAIDs: Nonsteroidal anti-inflammatory drugs; SD: Standard Deviation; IQR: Interquartile Range.

Table 1 (cont.): Baseline characteristics at inception date of metformin, sulphonylureas, rosiglitazone, pioglitazone and insulin

Characteristic	Metformin N = 121,637 N (%)	Sulphonylureas N = 76,863 N (%)	Rosiglitazone N = 22,636 N (%)	Pioglitazone N = 18,953 N (%)	Insulin N = 26,458 N (%)
History of:					
ACS	13,132 (10.8%)	9,017 (11.7%)	2,378 (10.5%)	1,989 (10.5%)	5,056 (19.1%)
Stroke	8,031 (6.6%)	5,944 (7.7%)	1,349 (6.0%)	1,172 (6.2%)	2,534 (9.6%)
Heart failure	5,294 (4.4%)	5,268 (6.9%)	831 (3.7%)	550 (2.9%)	2,765 (10.5%)
Stable angina	15,900 (13.1%)	11,009 (14.3%)	3,081 (13.6%)	2,511 (13.2%)	4,903 (18.5%)
Coronary heart disease	17,896 (14.7%)	12,698 (16.5%)	3,376 (14.9%)	2,706 (14.3%)	5,892 (22.3%)
Hyperlipidaemia	11,488 (9.4%)	6,685 (8.7%)	2,791 (12.3%)	2,572 (13.6%)	2,925 (11.1%)
Coronary revascularisation	4,583 (3.8%)	2,879 (3.7%)	902 (4.0%)	810 (4.3%)	1,597 (6.0%)
Hypertension	84,968 (69.9%)	53,462 (69.6%)	17,618 (77.8%)	15,424 (81.4%)	20,470 (77.4%)
Renal Impairment	6,630 (5.5%)	5,829 (7.6%)	1,173 (5.2%)	2,818 (14.9%)	3,501 (13.2%)
Peripheral vascular disease	5,716 (4.7%)	4,183 (5.4%)	1,255 (5.5%)	1,069 (5.6%)	2,305 (8.7%)
Recent prescribing of:					
Nitrates	11,664 (9.6%)	8,540 (11.1%)	2,158 (9.5%)	1,653 (8.7%)	4,308 (16.3%)
Beta blockers	29,206 (24.0%)	17,938 (23.3%)	5,636 (24.9%)	4,623 (24.4%)	7,065 (26.7%)
Calcium channel blockers	29,382 (24.2%)	18,532 (24.1%)	6,183 (27.3%)	5,681 (30.0%)	7,656 (28.9%)
Diuretics	41,234 (33.9%)	27,929 (36.3%)	7,651 (33.8%)	6,645 (35.1%)	10,917 (41.3%)
Antiplatelets	43,468 (35.7%)	27,971 (36.4%)	10,441 (46.1%)	10,049 (53.0%)	12,669 (47.9%)
ACE inhibitors/ angiotensin II receptor blockers	51,946 (42.7%)	31,870 (41.5%)	13,132 (58.0%)	12,371 (65.3%)	14,595 (55.2%)
Statins or fibrates	61,239 (50.3%)	35,295 (45.9%)	15,271 (67.5%)	14,934 (78.8%)	15,502 (58.6%)
NSAIDs	54,420 (44.7%)	34,424 (44.8%)	12,111 (53.5%)	11,071 (58.4%)	14,010 (53.0%)

ACE: angiotensin-converting-enzyme inhibitors; ACS: Acute Coronary Syndrome; BMI: Body Mass Index; HbA1c: Haemoglobin A1C test; NSAIDs: Nonsteroidal anti-inflammatory drugs; SD: Standard Deviation; IQR: Interquartile Range.

Table 2: Rates of outcomes (in GPRD) during past exposure compared to matched control cohort or to past metformin exposure

		Comparison with control cohort without diabetes		Comparison with past exposure of metformin	
Outcome	**Drug Class**	**Age, sex, calendar year Adjusted RR (95% CI)**	**Fully Adjusted RR (95% CI)**	**Age, sex, calendar year Adjusted RR (95% CI)**	**Fully Adjusted RR (95% CI)**
Death	Insulin	2.40 (2.20-2.61)	2.84 (2.60-3.11)	1.38 (1.26-1.50)	1.24 (1.12-1.37)
	Sulphonylureas	1.77 (1.72-1.82)	2.18 (2.11-2.26)	0.96 (0.89-1.03)	0.97 (0.90-1.04)
	Thiazolidinediones	1.77 (1.66-1.88)	3.04 (2.77-3.33)	0.81 (0.76-0.85)	0.84 (0.79-0.89)
	Metformin	2.01 (1.96-2.07)	2.23 (2.15-2.31)	Reference	Reference
	Controls	Reference	Reference	N/A	N/A
ACS	Insulin	2.00 (1.54-2.61)	1.68 (1.27-2.21)	1.00 (0.76-1.32)	1.04 (0.79-1.37)
	Sulphonylureas	2.38 (2.23-2.54)	1.63 (1.51-1.77)	1.23 (1.03-1.48)	1.29 (1.08-1.55)
	Thiazolidinediones	2.47 (2.17-2.81)	1.61 (1.31-1.97)	1.03 (0.91-.118)	1.05 (0.92-1.20)
	Metformin	2.48 (2.32-2.67)	1.55 (1.42-1.68)	Reference	Reference
	Controls	Reference	Reference	N/A	N/A
Stroke	Insulin	1.91 (1.45-2.52)	1.91 (1.43-2.56)	0.99 (0.75-1.32)	0.95 (0.71-1.26)
	Sulphonylureas	1.82 (1.68-1.97)	1.60 (1.45-1.76)	1.01 (0.82-1.25)	1.04 (0.85-1.29)
	Thiazolidinediones	1.90 (1.61-2.25)	1.89 (1.46-2.45)	0.91 (0.77-1.07)	0.92 (0.79-1.09)
	Metformin	2.05 (1.89-2.22)	1.72 (1.56-1.89)	Reference	Reference
	Controls	Reference	Reference	N/A	N/A
Congestive Heart Failure	Insulin	2.23 (1.74-2.87)	1.62 (1.25-2.10)	0.94 (0.73-1.22)	0.94 (0.73-1.22)
	Sulphonylureas	2.45 (2.30-2.62)	1.42 (1.31-1.54)	1.20 (0.99-1.44)	1.34 (1.12-1.62)
	Thiazolidinediones	3.23 (2.80-3.72)	1.75 (1.41-2.17)	1.12 (0.98-1.28)	1.13 (0.99-1.29)
	Metformin	2.97 (2.77-3.19)	1.50 (1.38-1.63)	Reference	Reference
	Controls	Reference	Reference	N/A	N/A

ACS: Acute Coronary Syndrome; CI: Confidence interval; GPRD: General Practice Research Database; RR: Relative Rate.

Table 3: Rates of outcomes during current exposure of pioglitazone and rosiglitazone

Outcome	Data source	Class	Number of cases	Incidence Rate[&]	Age, sex, calendar year Adjusted RR (95% CI)	Fully Adjusted RR (95% CI)
Death						
		Pioglitazone	487	1.7	Reference	Reference
	GPRD	Rosiglitazone	1274	2.0	1.15 (1.04-1.28)	1.20 (1.08-1.34)
		Pioglitazone	145	1.6	Reference	Reference
	ONS	Rosiglitazone	469	1.9	1.09 (0.90-1.32)	1.07 (0.89-1.29)
Acute Coronary Syndrome		Pioglitazone	219	0.9	Reference	Reference
	GPRD	Rosiglitazone	510	0.9	0.99 (0.84-1.16)	1.03 (0.87-1.21)
		Pioglitazone	67	0.8	Reference	Reference
	HES	Rosiglitazone	203	0.9	1.04 (0.79-1.37)	1.02 (0.77-1.35)
Stroke						
		Pioglitazone	108	0.4	Reference	Reference
	GPRD	Rosiglitazone	254	0.4	1.00 (0.80-1.26)	1.05 (0.83-1.32)
		Pioglitazone	31	0.4	Reference	Reference
	HES	Rosiglitazone	105	0.4	1.16 (0.77-1.74)	1.12 (0.75-1.68)
Congestive Heart Failure		Pioglitazone	208	0.7	Reference	Reference
	GPRD	Rosiglitazone	534	0.9	1.08 (0.92-1.27)	1.14 (0.97-1.34)
		Pioglitazone	34	0.4	Reference	Reference
	HES	Rosiglitazone	172	0.7	1.73 (1.19-2.50)	1.73 (1.19-2.51)

[&]Number of cases per 100 person years

ACS: Acute Coronary Syndrome; CI: Confidence interval; GPRD: General Practice Research Database; HES: Hospital Episode Statistics; ONS: Office for National Statistics; RR: Relative Rate.

Table 4: Rates of outcomes (in GPRD) during current exposure of rosiglitazone compared to pioglitazone stratified by age, co-prescribing of insulin and calendar time.

Outcome	Stratification	Age, sex, calendar year Adjusted RR (95% CI)	Fully Adjusted RR (95% CI)
Death	Age		
	<65	1.08 (0.82-1.41)	1.14 (0.87-1.49)
	≥65	1.17 (1.04-1.31)	1.22 (1.09-1.37)
	Co-prescribing insulin:		
	no	1.16 (1.05-1.30)	1.22 (1.09-1.35)
	yes	1.12 (0.63-1.97)	1.20 (0.68-2.13)
	Calendar year		
	<2007	1.15 (0.98-1.35)	1.20 (1.03-1.41)
	≥2007	1.13 (0.98-1.30)	1.19 (1.03-1.37)
Acute Coronary Syndrome	Age		
	<65	0.79 (0.61-1.03)	0.82 (0.64-1.07)
	≥65	1.12 (0.92-1.38)	1.17 (0.95-1.43)
	Co-prescribing insulin:		
	no	0.98 (0.83-1.15)	1.00 (0.85-1.18)
	yes	1.70 (0.71-4.08)	1.83 (0.76-4.42)
	<2007	1.04 (0.83-1.30)	1.06 (0.84-1.33)
	≥2007	0.90 (0.71-1.13)	0.94 (0.75-1.19)
Stroke	Age		
	<65	0.84 (0.52-1.36)	0.90 (0.56-1.46)
	≥65	1.05 (0.81-1.36)	1.09 (0.84-1.41)
	Co-prescribing insulin:		
	no	0.98 (0.78-1.23)	1.02 (0.81-1.29)
	yes	2.54 (0.49-13.16)	2.66 (0.51-13.84)
	<2007	1.33 (0.94-1.87)	1.40 (0.99-1.98)
	≥2007	0.74 (0.54-1.02)	0.76 (0.55-1.06)
Congestive Heart Failure	Age		
	<65	0.90 (0.61-1.34)	0.95 (0.64-1.42)
	≥65	1.12 (0.94-1.34)	1.19 (0.99-1.42)
	Co-prescribing insulin:		
	no	1.11 (0.94-1.31)	1.16 (0.98-1.37)
	yes	0.73 (0.31-1.75)	0.80 (0.33-1.91)
	<2007	1.04 (0.83-1.29)	1.10 (0.88-1.37)
	≥2007	1.13 (0.89-1.43)	1.20 (0.94-1.53)

CI: Confidence interval; GPRD: General Practice Research Database; RR: Relative Rate.

Table 5: Mortality rates and primary cause of death in rosiglitazone and pioglitazone current users

Primary cause of death (ICD-10 codes)	Rosiglitazone		Pioglitazone		Age, sex, calendar year Adjusted RR (95% CI)
	Number of cases	Incidence Rate[&]	Number of cases	Incidence Rate[&]	
All cause	469	1.9	145	1.63	1.09 (0.90 -1.32)
A00–B99: Certain infectious and parasitic diseases	7	0.03	7	0.08	0.30 (0.10 -0.87)
C00–D89: Neoplasms / Diseases of the blood and blood forming organs and certain disorders involving the immune mechanism	106	0.43	38	0.43	0.95 (0.65 -1.38)
E00–E90: Endocrine, nutritional and metabolic disorders	25	0.1	14	0.16	0.56 (0.29 -1.09)
F00–F99: Mental and behavioral disorders	3	0.01	1	0.01	1.30 (0.13 -12.77)
G00–G99: Diseases of the nervous system	1	0	3	0.03	0.11 (0.01 -1.05)
I00–I99: Diseases of the circulatory system	223	0.9	56	0.63	1.34 (1.00-1.80)
I20-I25: Ischemic heart disease	176	0.71	47	0.53	1.26 (0.91 -1.75)
I50: Heart failure	88	0.36	21	0.24	1.36 (0.84 -2.19)
J00–J99: Diseases of the respiratory system	55	0.22	10	0.11	1.91 (0.97 -3.77)
K00–K93: Diseases of the digestive system	25	0.1	11	0.12	0.78 (0.38 -1.59)
L00–L99: Diseases of the skin and subcutaneous tissue	4	0.02	0	0	
M00–M99: Diseases of the musculoskeletal system and connective tissue	1	0	1	0.01	0.30 (0.02 -5.00)
N00–N99: Diseases of the genitourinary system	8	0.03	1	0.01	2.43 (0.30 -19.57)
V01–Y98: External causes of morbidity and mortality	9	0.04	3	0.03	1.1 (0.29 -4.12)

[&]Number of cases per 100 person years

CI: Confidence interval; ICD-10: International Classification of Disease, version 10; RR: Relative Rate.

Table 6: Cumulative incidence (%) of outcomes (in GPRD) over one and three years in current rosiglitazone and pioglitazone users

			Year 1			Year 3		
Outcome	**Strata**		**Rosiglitazone**	**Pioglitazone**	**Excess risk**	**Rosiglitazone**	**Pioglitazone**	**Excess risk**
Death	All		1.85 (1.65-2.06)	1.48 (1.25-1.76)	0.36	5.80 (5.37-6.27)	4.46 (3.89-5.11)	1.35
	Age	40-49	0.21 (0.08-0.57)	0.12 (0.02-0.85)	0.09	0.82 (0.44-1.50)	0.56 (0.21-1.50)	0.25
		50-64	0.55 (0.40-0.75)	0.38 (0.23-0.64)	0.17	2.33 (1.91-2.84)	2.00 (1.44-2.76)	0.33
		65-74	1.77 (1.44-2.17)	1.67 (1.24-2.25)	0.09	6.15 (5.35-7.06)	4.19 (3.26-5.38)	1.96
		75-84	5.00 (4.20-5.95)	4.44 (3.43-5.74)	0.56	14.67 (12.95-16.59)	11.72 (9.46-14.48)	2.95
		85+	15.87 (12.54-19.99)	9.27 (5.83-14.57)	6.6	38.90 (32.34-46.26)	31.81 (22.41-43.89)	7.09
	Sex	female	1.97 (1.68-2.32)	1.71 (1.33-2.19)	0.26	6.46 (5.76-7.24)	5.39 (4.44-6.55)	1.07
		male	1.75 (1.51-2.03)	1.33 (1.05-1.68)	0.42	5.33 (4.79-5.93)	3.83 (3.17-4.63)	1.5
Death,	All		3.10 (2.83-3.39)	2.81 (2.46-3.21)	0.29	8.85 (8.28-9.45)	7.67 (6.88-8.55)	1.18
ACS,	Age	40-49	0.70 (0.42-1.17)	0.91 (0.48-1.73)	-0.21	1.69 (1.12-2.54)	1.87 (1.12-3.12)	-0.18
stroke,		50-64	1.40 (1.14-1.72)	1.73 (1.34-2.23)	-0.32	4.67 (4.05-5.39)	5.29 (4.32-6.47)	-0.62
heart failure		65-74	3.87 (3.32-4.51)	2.81 (2.19-3.61)	1.06	10.91 (9.78-12.17)	8.24 (6.76-10.02)	2.68
		75-84	7.42 (6.29-8.75)	6.86 (5.41-8.67)	0.57	20.64 (18.38-23.14)	16.43 (13.46-19.98)	4.21
		85+	19.91 (15.56-25.27)	16.25 (10.83-24.00)	3.66	45.53 (38.34-53.39)	36.60 (26.30-49.36)	8.93
	Sex	female	3.06 (2.67-3.51)	2.48 (2.00-3.08)	0.58	9.25 (8.38-10.20)	7.91 (6.67-9.36)	1.34
		male	3.12 (2.76-3.52)	3.04 (2.57-3.60)	0.07	8.54 (7.81-9.34)	7.50 (6.51-8.64)	1.04

ACS: Acute Coronary Syndrome; GPRD: General Practice Research Database.

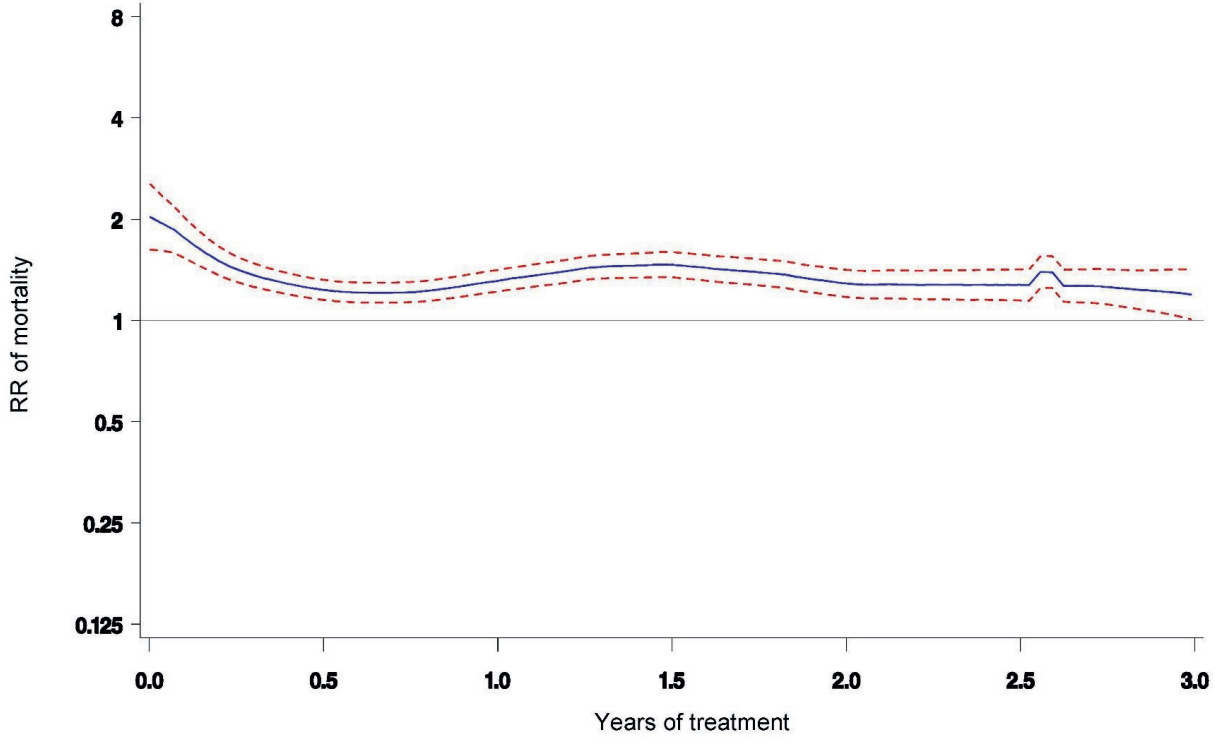


Figure 1: Smoothed crude RR of death over duration of treatment in rosiglitazone users compared to pioglitazone users

RR: Relative Rate.

Discussion

This study found that there was a substantive heterogeneity between the populations using different classes of various diabetes medications and that statistical adjustment with the measured risk factors only partially eliminated this bias. Comparable populations used rosiglitazone and pioglitazone. Higher risks for death (overall and due to cardiovascular disease) and heart failure were found for rosiglitazone compared to pioglitazone, with the risks of death highest shortly after starting treatment and disappearing after discontinuation of rosiglitazone.

Several studies have evaluated the cardiovascular safety of different classes of diabetes medications [7-32]. As an example, Tzoulaki *et al.* concluded that sulphonylureas had an unfavourable risk profile compared with metformin [25]. All of these studies relied on regression analysis to deal with confounding but none tested whether this statistical approach indeed minimised confounding. Although not routinely conducted, the formal analysis of bias has been recommended to be a critical part of an analysis [36]. We used two approaches for this bias analysis (one evaluating diabetes patients to controls and one comparing past exposure of different drugs). Both these analyses suggested that statistical adjustment only partially removed confounding due to the underlying disease. But there may be alternative explanations of these findings, namely that most diabetes drugs have cardiovascular side-effects or that various diabetes drugs have persistent effects (increasing the risks during past exposure). Although we can not exclude with certainty these alternative explanations, the presence of residual confounding due to underlying disease seems most likely as in diabetes mellitus drug exposure is defined by diabetes severity. As drug exposure is defined by diabetes severity, epidemiological comparisons between different classes of diabetes medications should not simply rely on statistical adjustment with a few risk factors but should evaluate the extent of bias in these comparisons.

Rosiglitazone and pioglitazone appeared to be used in the UK by comparable populations, as indicated by the lack of differences in cardiovascular risks during past exposure and general similarity in baseline risk factors. Our findings of higher risks with rosiglitazone could be explained by a relatively higher level of toxicity or by a lower level of benefit with rosiglitazone. Given the challenges in comparing between different classes of diabetes medication, our study could not evaluate whether the findings are explained by excess toxicity or by lesser benefit with rosiglitazone. As an example, pioglitazone is well recognised to cause heart failure and the decreased RR of heart failure with pioglitazone may only indicate that pioglitazone is less toxic than rosiglitazone. The results of this study are consistent with those reported by Graham *et al.* using US Medicare data [31], although the follow-up in the present study was considerably longer. Also, we did not censor at a non-endpoint hospitalisation, as this

could lead to differential loss to follow-up. A study by Wertz, that applied propensity matching to US claims data from insured employees, did not find any differences in the risk of death, MI and heart failure between rosiglitazone and pioglitazone [30]. Although this study appeared to be well conducted, the data concerned a restricted population with rates of death and MI lower than in the present study. Also, no results were provided without propensity score matching, which is a complex statistical technique. A recent meta-analysis of observational studies found a RR of mortality for rosiglitazone which were comparable to the present study, although it did report a small increased RR for myocardial infarction [37]. This meta-analysis did not adjust for data quality and consisted of pooling of studies with varying designs and study populations.

There is only indirect evidence from RCT comparing rosiglitazone and pioglitazone, as the large RCTs did not directly compare these drugs. But the RCTs do suggest differential effects. The RECORD study (comparing rosiglitazone with metformin or sulphonylurea to metformin plus sulphonylurea) found that the rate of cardiovascular death, MI and stroke was statistically comparable between the groups [38]. In contrast, the PROactive study found that this outcome was reduced with pioglitazone [39]. Similarly, the FDA meta-analysis found a trend for increased risks with rosiglitazone and decreased risks for pioglitazone (RRs of 2.14 and 0.54, respectively; p-values >0.05) [4]. The RRs for heart failure risks were also reported to be nominally higher with rosiglitazone than with pioglitazone [5]. These meta-analyses, lumping a set of heterogeneous studies, do not provide definitive evidence that rosiglitazone is inferior to pioglitazone. Observational studies, like the present one, are also limited by the potential for bias [39]. But despite the limitations of the present evidence base on the safety of rosiglitazone [40], an important consideration has been that there was limited evidence of a benefit of rosiglitazone on major clinical outcomes; the largest RCT (RECORD) did not show any beneficial effects of rosiglitazone [39]. The excess risks of cardiovascular outcomes in younger patients (under 65 years) was found to be small with rosiglitazone in this present study. Clearly, this finding of minimal adverse effects in younger patients needs be balanced by the lack of evidence of any beneficial effect of rosiglitazone in this group of patients.

There are significant challenges in interpreting the scientific evidence of the safety of medicines. Registration RCTs typically include a narrow set of patients recruited in specialised centres who are followed for a limited period of time. Often, these studies do not have the power to measure the effects on major clinical outcomes. Meta-analyses are now often conducted to overcome this limitation but the statistical lumping of heterogeneous RCTs is clearly not without limitations. The data from epidemiological studies often suffer from over-reliance on statistical adjustment and varying quality in data, design and study execution. Discrepant results within the

same database are indicative of the methodological challenges in epidemiological research [41]. Also, there is often incomplete evidence after approval of a medicine about the targeting of the treatment and who should receive it. Thus, there is a major need to expand our toolbox for obtaining evidence on the effects of medicines. One option could be large simple RCTs conducted within a research database, allowing the clinician to randomise between treatments and then following for outcomes using the routine electronic health records.

There are various limitations to this study. Information on confounders and underlying disease severity was limited in this study. Furthermore, our analyses provide only simplistic representations of the actual exposures to diabetes medications. Drug exposure in actual clinical practice often varies greatly, with many different drug combinations being used and patients switching over time between drugs and patients being non-compliant to treatment instructions. We did not evaluate this complexity in exposure and also relied on information of prescriptions rather than actual use.

In conclusion, the findings in this study support the presence of unmeasured confounding in the comparisons of cardiovascular outcomes between different classes of diabetes medications due to heterogeneity in use (as reflected by the substantive differences in rates of death during past exposure). Comparable populations used rosiglitazone and pioglitazone. Higher risks for death (overall and due to cardiovascular disease) and heart failure were found for rosiglitazone compared to pioglitazone, with the risks of death highest shortly after starting treatment and disappearing after discontinuation of rosiglitazone. These excess risks were largest in patients aged 65 years or older. This study supports the suspension of rosiglitazone by European regulatory authorities in September 2010.

References

1 Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med. 2007;356:2457-71.

2 Nissen SE, Wolski K. Rosiglitazone Revisited: An Updated Meta-analysis of Risk for Myocardial Infarction and Cardiovascular Mortality. Arch Intern Med. 2010;170(14):1191-201.

3 Lincoff AM, Wolski K, Nicholls SJ, Nissen SE. Pioglitazone and risk of cardiovascular events in patients with type 2 diabetes mellitus: a meta-analysis of randomized trials. JAMA. 2007;298:1180-8.

4 FDA website. Pioglitazone and rosiglitazone: cardiovascular safety meta-analyses. Available at: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/EndocrinologicandMetabolicDrugsAdvisoryCommittee/UCM224739.pdf [Accessed 14 January 2011]

5 Lago RM, Singh PP, Nesto RW. Congestive heart failure and cardiovascular death in patients with prediabetes and type 2 diabetes given thiazolidinediones: a meta-analysis of randomised clinical trials. Lancet. 2007;370:1129-36.

6 van Staa TP, Leufkens HG, Zhang B, Smeeth L. A comparison of cost effectiveness using data from randomized trials or actual clinical practice: selective cox-2 inhibitors as an example. PLoS Med. 2009;6:e1000194.

7 Rajagopalan R, Rosenson RS, Fernandes AW, Khan M, Murray FT. Association between congestive heart failure and hospitalization in patients with type 2 diabetes mellitus receiving treatment with insulin or pioglitazone: A retrospective data analysis. Clinical Therapeutics. 2004;26:1400-10.

8 Karter AJ, Ahmed AT, Liu J, Moffet HH, Parker MM. Pioglitazone initiation and subsequent hospitalization for congestive heart failure. Diabet Med. 2005;22, 986-93.

9 Masoudi FA, Inzucchi SE, Wang Y, Havranek EP, Foody JM, Krumholz HM. Thiazolidinediones, metformin, and outcomes in older patients with diabetes and heart failure: an observational study. Circulation. 2005;111:583-90.

10 Johannes CB, Koro CE, Quinn SG, Cutone JA, Seeger JD. The risk of coronary heart disease in type 2 diabetic patients exposed to thiazolidinediones compared to metformin and sulfonylurea therapy. Pharmacoepidemiol Drug Saf. 2007;16:504-12.

11 Gerrits CM, Bhattacharya M, Manthena S, Baran R, Perez A, Kupfer S. A comparison of pioglitazone and rosiglitazone for hospitalization for acute myocardial infarction in type 2 diabetes. Pharmacoepidemiol Drug Saf. 2007;16:1065-71.

12 Lipscombe LL, Gomes T, Levesque LE, Hux JE, Juurlink DN, Alter DA. Thiazolidinediones and cardiovascular outcomes in older patients with diabetes. JAMA. 2007;298:2634-43.

13 McAfee AT, Koro C, Landon J, Ziyadeh N, Walker AM. Coronary heart disease outcomes in patients receiving antidiabetic agents. Pharmacoepidemiol Drug Saf. 2007;16:711-25.

14 Koro CE, Fu Q, Stender M. An assessment of the effect of thiazolidinedione exposure on the risk of myocardial infarction in type 2 diabetic patients. Pharmacoepidemiol Drug Saf. 2008;17:989-96.

15 Margolis DJ, Hoffstad O, Strom BL. Association between serious ischemic cardiac outcomes and medications used to treat diabetes. Pharmacoepidemiol Drug Saf. 2008;17:753-9.

16 Walker AM, Koro CE, Landon J. Coronary heart disease outcomes in patients receiving antidiabetic agents in the PharMetrics database 2000-2007. Pharmacoepidemiol Drug Saf. 2008;17:760-8.
17 Winkelmayer WC, Setoguchi S, Levin R, Solomon DH. Comparison of cardiovascular outcomes in elderly patients with diabetes who initiated rosiglitazone vs pioglitazone therapy. Arch Intern Med. 2008;168:2368-75.
18 Dormuth CR, Maclure M, Carney G, Schneeweiss S, Bassett K, Wright JM. Rosiglitazone and myocardial infarction in patients previously prescribed metformin. PLOS One. 2009;4:e6080.
19 Dore, D D, Trivedi AN, Mor V, Lapane KL. Association between extent of thiazolidinedione exposure and risk of acute myocardial infarction. Pharmacotherapy. 2009;29:775-83.
20 Habib ZA, Tzogias L, Havstad SL, Wells K, Divine G, Lanfear DE, Tang J, Krajenta R, Pladevall M, Williams LK. Relationship between thiazolidinedione use and cardiovascular outcomes and all-cause mortality among patients with diabetes: A time-updated propensity analysis. Pharmacoepidemiol Drug Saf. 2009;18:437-47.
21 Hsiao FY, Huang WF, Wen YW, Chen PF, Kuo KN, Tsai YW. Thiazolidinediones and cardiovascular events in patients with type 2 diabetes mellitus: A retrospective cohort study of over 473,000 patients using the national health insurance database in Taiwan. Drug Saf. 2009;32:675-90.
22 Juurlink DN, Gomes T, Lipscombe LL, Austin PC, Hux JE, Mamdani MM. Adverse cardiovascular events during treatment with pioglitazone and rosiglitazone: Population based cohort study. BMJ. 2009;339:b2942.
23 Shaya FT, Lu Z, Sohn K, Weir MR. Thiazolidinediones and cardiovascular events in high-risk patients with type-2 diabetes mellitus: a comparison with other oral antidiabetic agents. P T. 2009;34:490-501.
24 Stockl KM, Le L, Zhang S, Harada AS. Risk of acute myocardial infarction in patients treated with thiazolidinediones or other antidiabetic medications. Pharmacoepidemiol Drug Saf. 2009;18:166-74.
25 Tzoulaki I, Molokhia M, Curcin V, Little MP, Millett CJ, Ng A, Hughes RI, Khunti K, Wilkins MR, Majeed A, Elliott P. Risk of cardiovascular disease and all cause mortality among patients with type 2 diabetes prescribed oral antidiabetes drugs: Retrospective cohort study using UK general practice research database. BMJ. 2009;339:b4731.
26 Vanasse A, Carpentier AC, Courteau J, Asghari S. Stroke and cardiovascular morbidity and mortality associated with rosiglitazone use in elderly diabetic patients. Diabetes & Vascular Disease Research. 2009;6:87-93.
27 Ziyadeh N, McAfee AT, Koro C, Landon J, Chan KA. The thiazolidinediones rosiglitazone and pioglitazone and the risk of coronary heart disease: A retrospective cohort study using a US health insurance database. Clinical Therapeutics. 2009;31:2665-77.
28 Bilik D, McEwen LN, Brown MB, Selby JV, Karter AJ, Marrero DG, Hsiao VC, Tseng CW, Mangione CM, Lasser NL, Crosson JC, Herman WH. Thiazolidinediones, cardiovascular disease and cardiovascular mortality: translating research into action for diabetes (TRIAD). Pharmacoepidemiol Drug Saf. 2010;19:715-21.
29 Wertz DA, Chang CL, Sarawate CA, Willey VJ, Cziraky MJ, Bohn RL. Risk of cardiovascular events and all-cause mortality in patients treated with thiazolidinediones in a managed-care population. Circ Cardiovasc Qual Outcomes. 2010;3:538-45
30 Graham DJ, Ouellet-Hellstrom R, MaCurdy TE, Ali F, Sholley C, Worrall C, Kelman JA. Risk of acute myocardial infarction, stroke, heart failure, and death in elderly patients with rosiglitazone or pioglitazone. JAMA. 2010;304:411-8.

31 Azoulay L, Schneider-Lindner V, Dell'aniello S, Filion KB, Suissa S. Thiazolidinediones and the risk of incident strokes in patients with type 2 diabetes: a nested case-control study. Pharmacoepidemiol Drug Saf. 2010;19:343-50.
32 Brownstein JS, Murphy SN, Goldfine AB, Grant RW, Sordo M, Gainer V, Collecchi JA, Dubey A, Nathan DM, Glaser JP, Kohane IS. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. Diabetes Care. 2010;33:526-31.
33 Parkinson J, Davis S, van Staa TP. The General Practice Research (GPRD) Database: Now and the Future. In Pharmacovigilance 2007. Editor R Mann.
34 Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69:4-14.
35 Ramlau-Hansen, H., Smoothing counting process intensities by means of kernel functional, Ann Statist. 1983;11:453-66.
36 Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. Springer Dordrecht Heidelberg London New York.
37 Loke YK, Kwok CS, Singh S. Comparative cardiovascular effects of thiazolidinediones: systematic review and meta-analysis of observational studies. BMJ. 2011;342:d1309.
38 Home PD, Pocock SJ, Beck-Nielsen H, Curtis PS, Gomis R, Hanefeld M, Jones NP, Komajda M, McMurray JJ; RECORD Study Team. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. Lancet. 2009;373:2125-35.
39 Dormandy JA, Charbonnel B, Eckland DJ, Erdmann E, Massi-Benedetti M, Moules IK, Skene AM, Tan MH, Lefèbvre PJ, Murray GD, Standl E, Wilcox RG, Wilhelmsen L, Betteridge J, Birkeland K, Golay A, Heine RJ, Korányi L, Laakso M, Mokán M, Norkus A, Pirags V, Podar T, Scheen A, Scherbaum W, Schernthaner G, Schmitz O, Skrha J, Smith U, Taton J; PROactive investigators. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. Lancet. 2005;366:1279-89.
40 GlaxoSmithKline website. GSK scientific overview of avandia data. Available at: http://www.gsk.com/avandia/GSK-Scientific-Overview-of-Avandia-Data.pdf [Accessed 18 October 2010]
41 de Vries F, de Vries CS, Cooper C, Leufkens HGM, van Staa TP. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the General Practice Research Database. Int J Epidemiol. 2006;35:1301-8.

Supporting information

Table S1: Cumulative incidence (%) of ACS, stroke and heart failure (in GPRD) over one and three years in current rosiglitazone and pioglitazone users

			Year 1			Year 3		
Outcome	**Strata**		**Rosiglitazone**	**Pioglitazone**	**Excess risk**	**Rosiglitazone**	**Pioglitazone**	**Excess risk**
ACS	All		1.06 (0.91-1.24)	0.94 (0.75-1.18)	0.12	2.43 (2.15-2.75)	2.41 (1.99-2.92)	0.02
	Age	40-49	0.36 (0.18-0.73)	0.71 (0.35-1.46)	-0.35	0.51 (0.27-0.97)	1.04 (0.54-2.00)	-0.52
		50-64	0.64 (0.47-0.86)	1.01 (0.72-1.41)	-0.37	1.66 (1.32-2.09)	2.46 (1.84-3.27)	-0.8
		65-74	1.55 (1.22-1.96)	0.72 (0.45-1.16)	0.82	3.17 (2.61-3.86)	1.75 (1.15-2.65)	1.43
		75-84	1.89 (1.37-2.61)	1.38 (0.83-2.31)	0.51	4.73 (3.67-6.08)	4.33 (2.91-6.40)	0.4
		85+	2.10 (0.93-4.72)	1.25 (0.30-5.18)	0.84	6.06 (3.31-10.98)	8.46 (3.10-21.94)	-2.4
	Sex	female	0.77 (0.59-1.01)	0.50 (0.31-0.81)	0.27	2.04 (1.65-2.52)	1.74 (1.19-2.56)	0.3
		male	1.29 (1.07-1.55)	1.26 (0.98-1.63)	0.03	2.74 (2.35-3.19)	2.89 (2.32-3.61)	-0.16
Stroke	All		0.55 (0.44-0.67)	0.38 (0.27-0.52)	0.17	1.23 (1.04-1.47)	0.99 (0.74-1.33)	0.24
	Age	40-49	0.04 (0.01-0.26)	0.06 (0.01-0.41)	-0.02	0.24 (0.07-0.83)	0.60 (0.21-1.72)	-0.36
		50-64	0.24 (0.15-0.39)	0.15 (0.07-0.35)	0.09	0.67 (0.46-0.97)	0.49 (0.26-0.94)	0.18
		65-74	0.54 (0.36-0.79)	0.47 (0.27-0.79)	0.07	1.44 (1.07-1.95)	1.41 (0.88-2.24)	0.03
		75-84	1.53 (1.10-2.14)	0.74 (0.39-1.40)	0.79	2.63 (1.93-3.60)	1.64 (0.87-3.06)	1.0
		85+	3.79 (2.13-6.70)	3.86 (1.68-8.76)	-0.07	8.17 (4.83-13.66)	3.86 (1.68-8.76)	4.31
	Sex	female	0.56 (0.41-0.77)	0.44 (0.28-0.70)	0.12	1.32 (1.02-1.71)	1.23 (0.82-1.87)	0.08
		male	0.54 (0.41-0.71)	0.33 (0.20-0.53)	0.21	1.18 (0.93-1.48)	0.83 (0.55-1.27)	0.34
Heart failure	All		0.82 (0.69-0.97)	0.77 (0.60-0.98)	0.05	2.51 (2.22-2.84)	2.13 (1.75-2.60)	0.38
	Age	40-49	0.07 (0.02-0.30)	0	0.07	0.23 (0.08-0.65)	0	0.23
		50-64	0.19 (0.11-0.33)	0.43 (0.26-0.72)	-0.24	0.93 (0.68-1.26)	1.09 (0.71-1.68)	-0.16
		65-74	0.98 (0.74-1.30)	0.71 (0.44-1.14)	0.27	3.05 (2.49-3.73)	2.84 (2.03-3.96)	0.21
		75-84	2.81 (2.19-3.60)	2.29 (1.57-3.34)	0.52	7.34 (6.04-8.90)	5.21 (3.81-7.11)	2.12
		85+	3.23 (1.73-6.01)	3.93 (1.72-8.89)	-0.7	12.87 (8.57-19.08)	6.55 (3.11-13.50)	6.32
	Sex	female	1.01 (0.80-1.27)	0.56 (0.36-0.88)	0.44	3.00 (2.53-3.56)	2.08 (1.50-2.88)	0.93
		male	0.67 (0.52-0.86)	0.91 (0.68-1.22)	-0.24	2.15 (1.81-2.56)	2.17 (1.69-2.79)	-0.02

ACS: Acute Coronary Syndrome; GPRD: General Practice Research Database

Chapter 4:

The impact on the study population and follow-up time when using linked data for observational research studies

Chapter 4.1

The impact of the choice of data source in record linkage studies estimating mortality in venous thromboembolism

Arlene M. Gallagher, Tim Williams, Hubert G.M. Leufkens, Frank de Vries

PLoS One. 2016;11(2):e0148349

Abstract

Linked electronic healthcare databases are increasingly being used in observational research. The objective of this study was to investigate the impact of the choice of data source in estimating mortality following VTE, with a secondary aim to investigate the influence of the denominator definition. We used the UK Clinical Practice Research Datalink (CPRD) to identify patients aged 18+ with venous thromboembolism (VTE). Multiple cohorts were identified in order to assess how mortality rates differed with a range of data sources. For each of the cohorts, incidence rates per 1,000 person years (/1000py) and relative rates (RRs) of all-cause mortality were calculated. The lowest mortality rate was found when only primary care data were used for both the exposure (VTE) and the outcome (death) (108.4/1000py). The highest mortality rate was found for patients diagnosed in secondary care (237.2/1000py). When linked primary and secondary care data were included for eligible patients and for the overlapping period of data collection, a mortality rate of 173.2/1000py was found. Sensitivity analyses varying the denominator definition provided a range of results (140.6-164.3/1000py). The relative rates of mortality by gender and age were comparable across all cohorts. Depending on the choice of data source, the population studied may be different. This may have substantial impact on the main findings, in particular on incidence rates of mortality following VTE.

Introduction

Electronic healthcare data bases are increasingly being used in observational research. For conditions diagnosed, treated and managed solely in primary care, electronic healthcare records (EHR) from general practice may be a good source of data for pharmacoepidemiology studies. However, some conditions have significant periods of management in secondary care or specialist centres and electronic data from one source may not capture all the events happening in another setting. In order to answer a specific research question, researchers must choose which data source(s) are most appropriate in order to ensure the findings are generalisable and that case misclassification can be minimised.

The impact of using different populations in EHR studies has not been evaluated. A previous publication investigated mortality following venous thromboembolism (VTE) using the combination of primary care, secondary care hospital episode statistics (HES) and mortality data from the Office for National Statistics (ONS) from the Clinical Practice Research Datalink (CPRD) [1]. Many of the symptoms of VTE present in primary care and a high level of validity of VTE recording in primary care has been reported previously [2,3]. More serious cases may first appear in secondary care or as a cause of death in the death certificate, without any record in the primary or secondary care records. General practitioners may not obtain the full information on the cause of death; feedback from secondary care can be poor, therefore death data can be non-coded or missing. Considering the range of recording options, this condition is an ideal candidate to use as an example of when there is a need to use multiple data sources to examine the patient journey in full.

Previous research has highlighted that case validity is improved by using more than one source of data [4], minimising misclassification of either the exposure or outcome, or both. The objective of this study was to investigate the impact of the choice of data source in estimating mortality following VTE, with a secondary aim to investigate the influence of the denominator definition. In order to do this we recreated an analysis of mortality following VTE using CPRD primary care data and linked HES and ONS mortality data, using the same version of the data sources and same study period [5]. We experimented with using the primary care data only and including available linked data.

Materials and methods

The primary care data from CPRD is a database of computerised medical records from across UK General Practice. Data are available from 1987 on over 13.6 million patients [6] in a dynamic cohort where patients can join/leave a GP practice over the course of follow-up. A systematic review of validation studies found that medical data in CPRD were generally of high validity [7]. The national Hospital Episodes Statistics (HES) data contain details of all admissions to National Health Service (NHS) hospitals in England from April 1997 [8]. The Office for National Statistics (ONS) mortality data contains the date and coded cause of death for the population of England and Wales from January 1998 [9]. HES and ONS mortality data are deterministically linked to the primary care data using a combination of identifiers including the patient's unique NHS number, gender, date of birth and postcode. For this study, each individual data source had a different period of coverage, the CPRD linkage was limited to consenting GP practices from England, and not all of the patients registered at a participating practice were eligible for each linkage (Figure 1).

We recreated an analysis of mortality following VTE using the same study period and data sources as in a previous publication [5]. An older version of the linked data was used covering the period up to the 30th October 2009, at which point 40% of CPRD practices had consented to take part in the linkage. VTE was defined as a composite of distal and proximal deep venous thrombosis (DVT) and pulmonary embolism (PE) from either primary or secondary care. In order to imitate making different choices, multiple cohorts were identified with the data sources available (Figure 2). In all cohorts, patients were aged 18 years or older with a diagnosis of VTE (index date) on or after 01/01/1995.

4.1

Four cohorts were identified in order to assess the impact of the choice of data source:

1. Primary care cohort; data on exposure (VTE) and outcome (death) from primary care only, for the whole study period 1995-2009
2. Linked exposure cohort; including linked exposure data from HES, for eligible patients and the overlapping study period 1997-2009
3. Linked exposure and outcome cohort; including linked exposure data from HES and linked outcome data from ONS, for eligible patients and the overlapping the study period 1998-2009
4. Secondary care cohort; data on exposure from HES only and outcome data from primary care, for the overlapping study period 1997-2009

Patients were categorised by their prior VTE risk based on modifiable (fractures, surgery, trauma) or unmodifiable (cancer, congestive heart failure, varicose veins) factors recorded before the index date. Patients were followed from the index date

until censoring, based upon the earliest of the end of data collection, patient's death, or transfer out of the practice, whichever date came first. The date of death was identified from the primary care data or linked ONS mortality data, where available. The mortality incidence rate for each of the cohorts was calculated per 1,000 person years (/1000py). Survival analyses using Cox proportional hazards regression was used to estimate the relative hazard rates (RRs) for all-cause mortality over time. RRs were calculated by gender and age using those aged 18-39 as the reference. The statistically adjusted model included age, gender, lifestyle information (BMI, smoking status, alcohol use) and VTE risk category.

Sensitivity analyses evaluated the influence of the denominator definition by considering individual eligibility and data coverage periods in two further cohorts; patients with VTE from either primary or secondary care over the whole study period (cohort A: all data cohort), and patients from consenting practices whether or not they were individually eligible (cohort B: consenting practices).

The CPRD Group has obtained ethical approval from a National Research Ethics Service Committee (NRES) for all purely observational research using anonymised CPRD data; namely, studies which do not include patient involvement (which is the vast majority of CPRD studies). Individual patient consent is not required for observational studies using anonymised CPRD data. Individual studies must be granted approval by the Independent Scientific Advisory Committee for MHRA database research (ISAC). This study was approved under protocol number 14/024.

Results

A total of 46,332 patients were identified with a record of VTE from either primary or secondary care between 1995 and 2009. The largest cohort was of those diagnosed in primary care (cohort 1: N=36,216, 71%), of which 3,320 (9%) were diagnosed before April 1997 when linked HES data became available and 6,539 (18%) were registered at practices in Wales, Scotland or Northern Ireland. Over 70% of patients had no obvious risk factor of VTE. In the HES data, 13,404 patients were identified with a diagnosis of VTE between 1997 and 2009 (cohort 4), of which 3,288 (25%) had a matching record in primary care on the same date. After pooling the primary care and HES patients and restricting to those eligible for HES linkage and to the overlapping coverage period (1997-2009), the linked exposure cohort included 23,720 cases (cohort 2). The average follow-up period for cohort 2 was shorter than cohort 1 by 0.5 years (3.54 vs. 4.08 years) and patients were more likely to have a modifiable risk factor for VTE (28.7% vs. 21.3%). The linked exposure and outcome cohort was created by restricting to those eligible for ONS linkage and to the overlapping coverage period (1998-2009), including 22,639 cases (cohort 3). The average follow-up period for cohort 3 was shorter than for both cohorts 1 and 2 at 3.39 years. The broadest cohort included all patients regardless of eligibility for linkage or whether the practice had consented (cohort A: all data cohort). This cohort included 46,296 patients of which 5,365 patients (11.6%) were diagnosed before mortality statistics were available from the ONS. The cohort of patients from consenting practices with a diagnosis in either primary care or HES included 26,320 patients (cohort B: consenting practices). A larger proportion of these were diagnosed in primary care (49% vs 43% in the other linked cohorts). Table 1 compares patient characteristics between all cohorts and shows that the age, gender, smoking and BMI profile was similar for all cohorts. The mean age was 55-56 years and 55% of patients were female. Over half of the patients were overweight or obese at the time of diagnosis and one fifth were smokers.

Table 2 shows that the mortality rates over the 2 years following a VTE diagnosis were lowest in the primary care cohort (108.4/1000py) and highest for patients diagnosed in secondary care (237.2/1000py). Including eligible patients diagnosed in either setting (linked exposure cohort) produced a mortality rate between the two (169.9/1000py); adding mortality outcome data from ONS (linked exposure and outcome cohort) resulted in an increase in the mortality rate (173.2/1000py). Restricting to practices participating in the linkage, but ignoring individual eligibility resulted in a rate lower rate than the linked exposure and outcome cohort (164.3 vs 173.2/1000py). The largest number of deaths (9,701) was attributed to the all data cohort; however the mortality rate was the lowest of all the cohorts that included linked data (140.6/1000py). Figure 3 compares all cohorts over the two years.

Table 3 shows that 85% of all fatal cases were in patients aged 60 or over. Older patients, especially the elderly, had an increased risk of mortality after adjusting for gender, lifestyle information (BMI, smoking status, alcohol use) and VTE risk category. A larger number of deaths were identified in women, however after adjustment, the mortality risk was found to be higher in men. The relative rate of mortality by gender and age were comparable across the cohorts although there was a lower mortality rate in the reference group for age (18-39) in the primary care cohort than all of the others.

Table 1: Baseline information on VTE patients at diagnosis.

Characteristic	1: Primary care cohort N=36,216	2: Linked exposure cohort N=23,720	3: Linked exposure and outcome cohort N=22,639	4: Secondary care cohort N=13,404	A: All Data cohort N=46,296	B: Consenting practice cohort N=26,320
Mean follow-up time (SD)	4.08 (3.56)	3.54 (3.24)	3.39 (3.08)	3.18 (3.20)	3.85 (3.50)	3.79 (3.55)
VTE diagnosis point, N (%)						
Primary care	36,216 (100%)	10,316 (43.5%)	9,814 (43.3%)	-	32,906 (71.1%)	12,930 (49.1%)
Hospital	-	10,116 (42.6%)	9,668 (42.7%)	13,404 (100%)	10,101 (21.8%)	10,101 (38.4%)
Both	-	3,288 (13.9%)	3,157 (13.9%)	-	3,289 (7.1%)	3,289 (12.5%)
Year of diagnosis, N (%)						
1995	1,394 (3.8%)	-	-	-	1,394 (3.0%)	691 (2.6%)
1996	1,490 (4.1%)	-	-	-	1,490 (3.2%)	744 (2.8%)
1997	1,600 (4.4%)	757 (3.2%)	-	413 (3.1%)	1,920 (4.1%)	1,078 (4.1%)
1998-2009	31,732 (87.6%)	22,963 (96.8%)	22,639 (100%)	12,991 (96.9%)	41,492 (89.6%)	23,807 (90.5%)
Age						
Mean (sd)	54.8 (17.9)	56.5 (17.6)	56.5 (17.6)	58.0 (17.4)	55.6 (17.8)	56.2 (17.7)
18-39	3,451 (9.5%)	1,895 (8.0%)	1,815 (8.0%)	935 (7.0%)	4,127 (8.9%)	2,199 (8.4%)
40-59	9,231 (25.5%)	5,444 (23.0%)	5,197 (23.0%)	2,746 (20.5%)	11,196 (24.2%)	6,129 (23.3%)
60-79	16,515 (45.6%)	11,064 (46.6%)	10,536 (46.5%)	6,362 (47.5%)	21,266 (45.9%)	12,198 (46.3%)
80+	7,019 (19.4%)	5,317 (22.4%)	5,091 (22.5%)	3,361 (25.1%)	9,707 (21.0%)	5,794 (22.0%)
Gender, N (%)						
Women	19,974 (55.2%)	13,244 (55.8%)	12,620 (55.7%)	7,366 (55.0%)	25,559 (55.2%)	14,664 (55.7%)
Men	16,242 (44.8%)	10,476 (44.2%)	10,019 (44.3%)	6,038 (45.0%)	20,737 (44.8%)	11,656 (44.3%)

Table 1 (cont.): Baseline information on VTE patients at diagnosis.

Characteristic	1: Primary care cohort N=36,216	2: Linked exposure cohort N=23,720	3: Linked exposure and outcome cohort N=22,639	4: Secondary care cohort N=13,404	A: All Data cohort N=46,296	B: Consenting practice cohort N=26,320
BMI, N (%)						
Underweight	1,630 (4.5%)	1,172 (4.9%)	1,132 (5.0%)	692 (5.2%)	2,179 (4.7%)	1,274 (4.8%)
Normal	8,539 (23.6%)	5,770 (24.3%)	5,523 (24.4%)	3,267 (24.4%)	11,032 (23.8%)	6,387 (24.3%)
Overweight	11,506 (31.8%)	7,380 (31.1%)	7,030 (31.1%)	4,082 (30.5%)	14,473 (31.3%)	8,203 (31.2%)
Obese	9,771 (27.0%)	6,185 (26.1%)	5,960 (26.3%)	3,288 (24.5%)	12,162 (26.3%)	6,741 (25.6%)
Unknown BMI	4,770 (13.2%)	3,213 (13.5%)	2,994 (13.2%)	2,075 (15.5%)	6,450 (13.9%)	3,715 (14.1%)
Smoking Status, N (%)						
Non Smoker	16,541 (45.7%)	10,667 (45.0%)	10,101 (44.6%)	5,928 (44.2%)	20,960 (45.3%)	11,958 (45.4%)
Ex-Smoker	9,187 (25.4%)	6,615 (27.9%)	6,468 (28.6%)	3,804 (28.4%)	12,024 (26.0%)	6,983 (26.5%)
Smoker	7,151 (19.7%)	4,412 (18.6%)	4,225 (18.7%)	2,440 (18.2%)	8,970 (19.4%)	4,953 (18.8%)
Unknown Smoking Status	3,337 (9.2%)	2,026 (8.5%)	1,845 (8.1%)	1,232 (9.2%)	4,342 (9.4%)	2,426 (9.2%)
Alcohol Drinking Status, N (%)						
Non Drinker	5,056 (14.0%)	3,010 (12.7%)	2,838 (12.5%)	1,717 (12.8%)	6,387 (13.8%)	3,405 (12.9%)
Ex-Drinker	2,045 (5.6%)	1,478 (6.2%)	1,447 (6.4%)	829 (6.2%)	2,672 (5.8%)	1,546 (5.9%)
Drinker	22,634 (62.5%)	15,157 (63.9%)	14,546 (64.3%)	8,451 (63.0%)	28,852 (62.3%)	16,659 (63.3%)
Unknown Drinking Status	6,481 (17.9%)	4,075 (17.2%)	3,808 (16.8%)	2,407 (18.0%)	8,385 (18.1%)	4,710 (17.9%)

Table 1 (cont.): Baseline information on VTE patients at diagnosis.

Characteristic	1: Primary care cohort N=36,216	2: Linked exposure cohort N=23,720	3: Linked exposure and outcome cohort N=22,639	4: Secondary care cohort N=13,404	A: All Data cohort N=46,296	B: Consenting practice cohort N=26,320
Risk category for VTE prior to index, N (%)						
Modifiable strong[α]	2,923 (8.1%)	3,205 (13.5%)	3,067 (13.5%)	1,825 (13.6%)	4,292 (9.3%)	3,326 (12.6%)
Modifiable moderate/low[β]	4,792 (13.2%)	3,604 (15.2%)	3,424 (15.1%)	2,117 (15.8%)	6,442 (13.9%)	3,937 (15.0%)
Unmodifiable[γ]	2,231 (6.2%)	1,647 (6.9%)	1,570 (6.9%)	1,023 (7.6%)	3,064 (6.6%)	1,833 (7.0%)
No obvious risk factor	26,270 (72.5%)	15,264 (64.4%)	14,578 (64.4%)	8,439 (63.0%)	32,498 (70.2%)	17,224 (65.4%)
Region, N (%)						
North East	607 (1.7%)	276 (1.2%)	261 (1.2%)	171 (1.3%)	731 (1.6%)	308 (1.2%)
North West	5,280 (14.6%)	4,406 (18.6%)	4,157 (18.4%)	2,615 (19.5%)	7,376 (15.9%)	4,863 (18.5%)
Yorkshire & The Humber	1,953 (5.4%)	1,318 (5.6%)	1,266 (5.6%)	819 (6.1%)	2,600 (5.6%)	1,446 (5.5%)
East Midlands	1,999 (5.5%)	885 (3.7%)	848 (3.7%)	518 (3.9%)	2,393 (5.2%)	996 (3.8%)
West Midlands	3,471 (9.6%)	3,097 (13.1%)	2,965 (13.1%)	1,706 (12.7%)	4,753 (10.3%)	3,350 (12.7%)
East of England	3,668 (10.1%)	3,231 (13.6%)	3,047 (13.5%)	1,728 (12.9%)	4,980 (10.8%)	3,641 (13.8%)
South West	2,874 (7.9%)	3,245 (13.7%)	3,081 (13.6%)	1,976 (14.7%)	4,368 (9.4%)	3,613 (13.7%)
South Central	3,884 (10.7%)	2,636 (11.1%)	2,543 (11.2%)	1,462 (10.9%)	4,792 (10.4%)	2,908 (11.0%)
London	3,111 (8.6%)	2,177 (9.2%)	2,124 (9.4%)	1,062 (7.9%)	3,955 (8.5%)	2,425 (9.2%)
South East Coast	2,830 (7.8%)	2,449 (10.3%)	2,347 (10.4%)	1,347 (10.0%)	3,809 (8.2%)	2,770 (10.5%)
Northern Ireland, Scotland or Wales	6,539 (18.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6,539 (14.1%)	0 (0.0%)

[α] Modifiable (transient) strong risk factor for VTE: hip or leg fracture, hip or knee replacement, major surgery (pelvis, lower leg, knee, feet, veins, arteries, spinal cord, and unspecified region), major trauma (head, neck, spinal cord, trunk, pelvis, lower leg, knee, feet, and unspecified region).

[β] Modifiable (transient) moderate/low risk factor for VTE: major surgery (head, neck, trunk, and arms/hands), major trauma (arms/hands), pneumonia, chronic obstructive pulmonary disease, hormone therapy, or oral contraceptives.

[γ] Unmodifiable risk factor for VTE: active cancer, congestive heart failure, varicose veins.

BMI: body mass index; SD: Standard Deviation; VTE: venous thromboembolism.

Table 2: Mortality rates following VTE, by cohort

Characteristic	N	Source of mortality data	Failures (deaths)	Person-time (years)	Incidence of Mortality per 1,000 person-years
1: Primary care cohort	36,216	primary care	6,112	56362	108.4
2: Linked exposure cohort	23,720	primary care	5,771	33961	169.9
3: Linked exposure and linked outcome cohort	22,639	primary care or ONS	5,569	32149	173.2
4: Secondary care cohort	13,404	primary care	4,186	17651	237.2
A: All data cohort	46,296	primary care or ONS	9,701	69001	140.6
B: Consenting practice cohort	26,320	primary care or ONS	6,264	38125	164.3

ONS: Office for National Statistics; VTE: venous thromboembolism.

Table 3: Relative hazard rates (RRs) for all-cause mortality following VTE, by cohort

	1: Primary care cohort N=36,216			2: Linked exposure cohort N=23,720			3: Linked exposure and outcome cohort N=22,639			4: Secondary care cohort N=13,404		
	N deaths	IR	Adjusted RR §	N deaths	IR	Adjusted RR §	N deaths	IR	Adjusted RR §	N deaths	IR	Adjusted RR §
All	6,112	108.4	-	5,771	169.9	-	5,569	173.2	-	4,186	237.2	-
Age												
18-39	89	15.5	Reference	99	32.4	Reference	95	32.7	Reference	72	49.7	Reference
40-59	859	56.3	3.99 (3.21 - 4.97)*	710	81.2	2.67 (2.16 - 3.29)*	684	82.4	2.70 (2.17 - 3.34)*	502	119.2	2.54 (1.98 - 3.25)*
60-79	3,095	120.4	8.54 (6.91 - 10.56)*	2,837	177.2	5.63 (4.60 - 6.89)*	2,735	181	5.72 (4.65 - 7.03)*	2,036	237.4	4.75 (3.75 - 6.02)*
80+	2,069	213.9	12.19 (9.83 - 15.10)*	2,125	345.4	8.76 (7.14 - 10.74)*	2,055	352.1	8.87 (7.21 - 10.93)*	1,576	461.9	7.24 (5.70 - 9.20)*
Gender												
Men	2,744	108.8		2,565	170.9	Reference	2,469	173.4	Reference	1,867	231.7	Reference
Women	3,368	108.1	0.90 (0.85 - 0.95)*	3,206	169.2	0.88 (0.83 - 0.93)*	3,100	173.1	0.89 (0.84 - 0.94)*	2,319	241.8	0.92 (0.86 - 0.98)*

Table 3 (cont.): Relative hazard rates (RRs) for all-cause mortality following VTE, by cohort

	A: All data cohort N=46,296			B: Consenting practice cohort N=26,320		
	N deaths	IR	Adjusted RR §	N deaths	IR	Adjusted RR §
All	9,701	140.6	-	6,264	164.3	-
Age						
18-39	154	22.8	Reference	107	30.0	Reference
40-59	1,283	70.5	3.32 (2.81 - 3.93)*	801	81.1	2.87 (2.34 - 3.51)*
60-79	4,795	150.5	6.92 (5.89 - 8.13)*	3,065	171.7	5.88 (4.84 - 7.14)*
80+	3,469	285.1	10.45 (8.88 - 12.31)*	2,291	335.9	9.18 (7.54 - 11.16)*
Gender						
Men	4,315	139.4		2,784	164.7	
Women	5,386	141.6	0.90 (0.87 - 0.94)*	3,480	164	0.88 (0.84 - 0.93)*

§ RR (95 CI), adjusted for age, gender, lifestyle information (BMI, smoking status, alcohol use) and VTE risk category
* P<0.05

CI: Confidence interval; IR: Incidence rate per 1,000 person-years; RR: Relative hazard rate; VTE: venous thromboembolism.

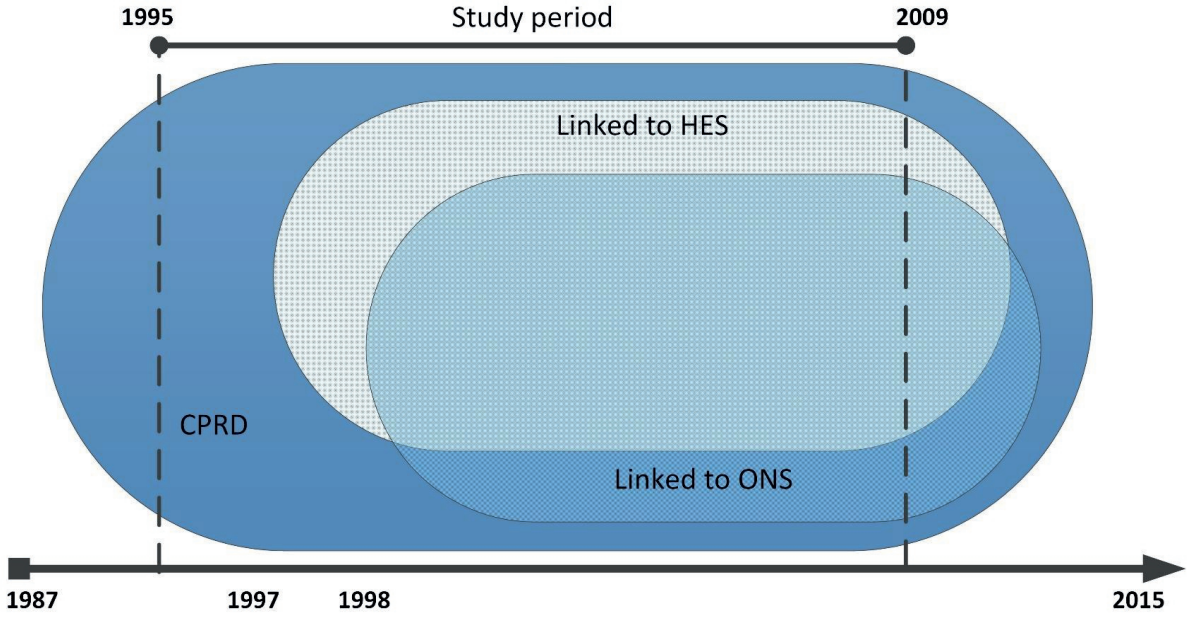


Figure 1: Availability of data within the study period.

CPRD: Clinical Practice Research Datalink; HES: Hospital Episode Statistics; ONS: Office for National Statistics

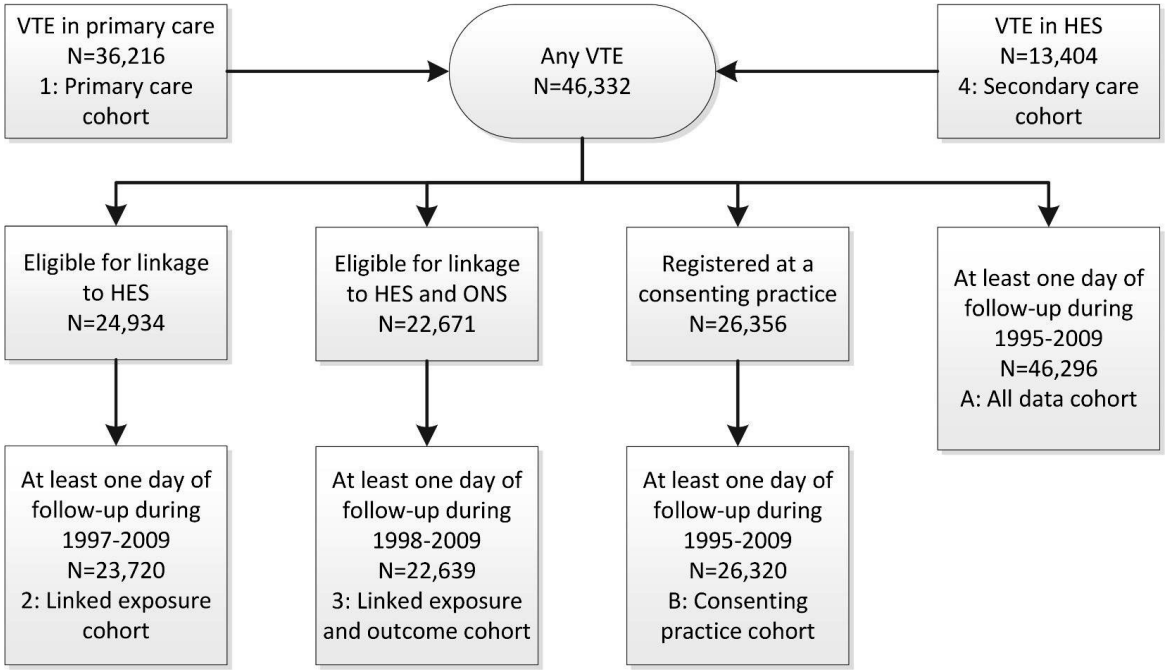


Figure 2: Flow chart of eligibility.

HES: Hospital Episode Statistics; VTE: venous thromboembolism.

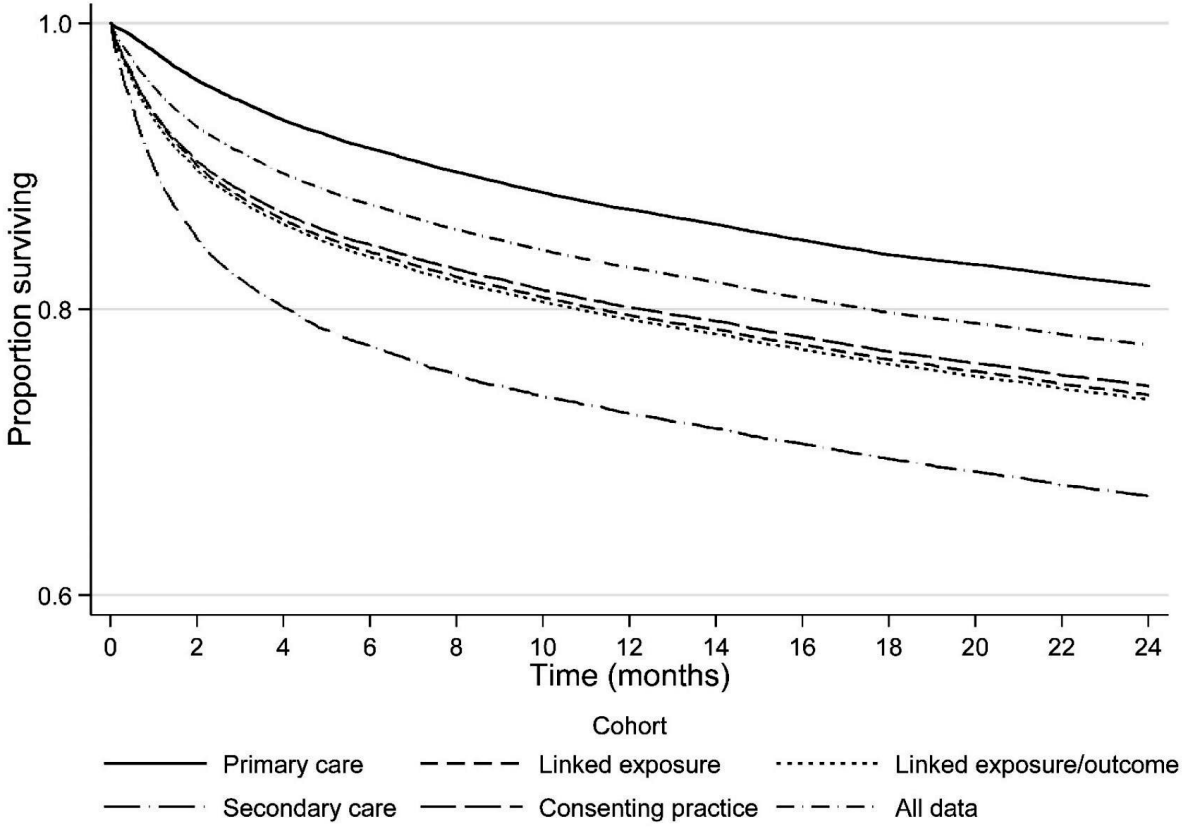


Figure 3: Mortality rate following VTE over time, by cohort

Discussion

We looked at the impact of the choice of data source in estimating mortality following VTE and found considerable differences in the observed mortality rate across cohorts built using different sources, despite comparable relative rates by gender and age. All of the cohorts we examined were similar in terms of age, gender and lifestyle information; however the mortality rate was two times higher for patients diagnosed in secondary care data compared to primary care. Using linked data sources provided estimates between the two; however the choice of denominator in linked populations also affected the estimates reported, with lower rates seen when individual eligibility was not considered and the whole study period was used. Figure 3 shows the variation in rates across all of the cohorts over two years, highlighting that the difference can be seen immediately following diagnosis. This suggests that the cohorts were representing different populations.

We repeated a previously published study, modifying the study design to identify four cohorts that could feasibly be used to analyse mortality following VTE. We estimated the incidence rate for each cohort and the relative risk of mortality comparing those aged 18-39 to other age categories and comparing men to women. The primary care cohort represented the patient set and analyses that would be performed if only general practice records were available. Both the exposure (VTE) and outcome (death) were identified from primary care records and data were available for the full study period 1995 - 2009. The mortality rate (108.4/1000py) was comparable to a previously published study (97.8/1000py) [10], however there are potential limitations on the suitability of studying VTE in a primary care cohort and the availability and accuracy of the mortality information.

A pulmonary embolism is a potentially life-threatening condition with common symptoms including chest pains and a shortness of breath. Patients with such symptoms are more likely to attend the hospital and one third of patients with symptomatic VTE manifest pulmonary embolism [11]. In analysing only patients diagnosed in secondary care, we found a mortality rate of more than double that in the primary care cohort (237.2 vs. 108.4/1000py). The overlap between cohorts was relatively small (7%), suggesting they were very different patient populations. This study is in line with previous research highlighting how data sources do not always concur [12,13,14]. Movig *et al.* (2002) compared clinical coded diagnoses from hospital with laboratory data and found a large number of patients (approximately 87%) identified via laboratory data did not have an associated coded diagnosis [14]. A study based on coded data may return a differing result to one based on laboratory results much like we have found when comparing patients diagnosed in primary versus secondary care.

Previous research has highlighted that case validity is improved by using a combination of linked data sources. Herrett *et al.* (2013) used four linked data sources to analyse acute myocardial infarction and found that each individual data source missed a substantial proportion (25-50%) of MI events, suggesting that failure to used linked data is likely to lead to biased estimates [4]. A more representative cohort of VTE cases may therefore include exposures from both primary and secondary care. The linked exposure cohort included diagnoses from both data sources for the overlapping coverage period (1997-2009). The mortality rate (169.9/1000py) was halfway between what was seen in the primary care (108.4/1000py) and secondary care (237.2/1000py) cohorts. This may have addressed potential misclassification of the exposure, however misclassification of the outcome may persist through the use of primary care data for the outcome.

Although the gatekeepers to healthcare access in the UK, holding the longitudinal record for patients, the GP may not always be informed of a patient's death in good time, and in some cases not at all. Within the GP record, some patients may have multiple records relating to death on different dates without clarity on which is correct. The ONS mortality dataset includes information on the date and cause of death from civil registration records. All deaths in England and Wales have to be registered and usually within five days. With just one record per person and up-to date recording, this provides the gold-standard data source for mortality information which could reduce the level of misclassification of the outcome. The addition of ONS mortality data with the relevant coverage period (1998-2009), resulted in a smaller cohort (N=22,639) with reduced person time (32,149) and an increased rate of 173.2/1000py. This cohort may have minimised misclassification, however it is not clear whether combining primary and secondary care data sources evaluates two different populations, where one comprising VTE cases presenting at hospital may be more severe.

Studies analysing linked data sources are frequently used in regulatory decision making, however previous publications have applied differing approaches in the definition of the denominator. With this in mind, we looked at the influence of not considering individual eligibility and data coverage periods. Linked data are not available for all patients in CPRD. Consent is given at the practice level and identifiers are sent to a trusted third party for linkage. Some registered patients may not be eligible for linkage if they have incomplete identifier information, withdraw consent or live outside of England. The consenting practice cohort included all patients from these practices regardless of eligibility. The cohort was necessarily larger than the linked cohorts, with more patients diagnosed in primary care (49%) and longer follow-up, presenting a lower mortality rate. The all data cohort explored eligibility further by including all patients and incorporating linked data where it was available; resembling the patient set and analyses reported in some previous

publications [15,16,17]. This cohort primarily included patients from primary care (71%), and included longer follow-up, presenting a mortality rate 20% lower than the linked exposure and outcome cohort. This lower rate is not unexpected since 11.6% of the patients in this cohort were diagnosed before the start of coverage of the ONS data creating an immortal time bias. These sensitivity analyses highlight the importance of considering the eligibility in the methodology using linked data alongside the influence of the two different populations.

There were certain limitations to this study. We did not confirm the VTE record or validate the death information; further analyses of the cause of death could highlight that some of deaths were unrelated to VTE. In addition, cases fatal on index date and those where the VTE was diagnosed as part of a post-mortem may not have been included in the study. We did not have access to data from outpatients or emergency visits. The secondary care cohort is limited to those linked to the primary care record rather than all patients presenting to hospital with VTE. It is important to note that the data sources are collected for use in routine clinical practice, for audit purposes or for the purpose of reporting national statistics rather than for research. The linked population was limited to England only, however a previous comparison of those included in the linkage programme to the whole CPRD population found no significant differences in patient characteristics [1] and similarly we found no substantial differences between the cohorts in this study (Table 1).

The size of the available cohort from CPRD makes the primary care EHR a viable choice for studying VTE however given the likelihood that more severe patients will present in secondary care, a linked population including patients presenting in secondary care may be a more appropriate source for analysis. Consideration of the influence of the denominator definition is important in studies of linked data as restricting the study population to those eligible for linkage does reduce the follow-up time, which is likely to have a large influence on small cohorts or those with rare outcomes. Future studies should consider the impact of the data source on case validity alongside the benefits of using linked data, together with the denominator definition. The choice of data source in this study demonstrated a substantial impact on the absolute rates of mortality following VTE. We used mortality following VTE as a learning case and expect these findings also may have an impact on other disease/outcome pairs.

References

1. Gallagher AM, Puri S, van Staa TP. Linkage of the General Practice Research Database (GPRD) with Other Data Sources. Pharmacoepidemiol Drug Saf. 2011;20:S230.
2. Lawrenson R, Todd JC, Leydon GM, Williams TJ, Farmer RD. Validation of the diagnosis of venous thromboembolism in general practice database studies. Br J Clin Pharmacol. 2000;49(6):591-6.
3. Huerta C, Johansson S, Wallander MA, García Rodríguez LA. Risk factors and short-term mortality of venous thromboembolism diagnosed in the primary care setting in the United Kingdom. Arch Intern Med. 2007;167(9):935-43.
4. Herrett E, Shah AD, Boggon R, *et al*. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013; 346:f2350.
5. Gallagher AM, de Vries F, Plumb JM, *et al*. Quality of INR control and outcomes following venous thromboembolism. Clin Appl Thromb Hemost. 2012;18(4):370-8.
6. Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36.
7. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69(1):4-14.
8. Health & Social Care Information Centre (HSCIC). Hospital Episode Statistics. Available at: http://www.hscic.gov.uk/hes [Accessed 15 June 2015].
9. HSCIC. Mortality data from the Office for National Statistics. Available at: http://www.hscic.gov.uk/onsmortality [Accessed 15 June 2015].
10. Huerta C, Johansson S, Wallander MA, Rodríguez LA. Risk of myocardial infarction and overall mortality in survivors of venous thromboembolism. Thromb J. 2008;6:10.
11. White RH. The epidemiology of venous thromboembolism. Circulation. 2003;107(23 Suppl 1):I4-8. Review.
12. Robertson LM, Denadai L, Black C, *et al*. Is routine hospital episode data sufficient for identifying individuals with chronic kidney disease? A comparison study with laboratory data. Health Informatics J. 2016;22(2):383-96
13. Lo Re V 3rd, Haynes K, Goldberg D, *et al*. Validity of diagnostic codes to identify cases of severe acute liver injury in the US Food and Drug Administration's Mini-Sentinel Distributed Database. Pharmacoepidemiol Drug Saf. 2013;22(8):861-72.
14. Movig KL, Leufkens HG, Lenderink AW, Egberts AC. Validity of hospital discharge International Classification of Diseases (ICD) codes for identifying patients with hyponatremia. J Clin Epidemiol. 2003;56(6):530-5.
15. Bazelier MT, van Staa T, Uitdehaag BM, *et al*. The risk of fracture in patients with multiple sclerosis: the UK general practice research database. J Bone Miner Res. 2011;26(9):2271-9.
16. Bazelier MT, van Staa TP, Uitdehaag BM, *et al*. A simple score for estimating the long-term risk of fracture in patients with multiple sclerosis. Neurology. 2012;79(9):922-8.
17. Bazelier MT, Mueller-Schotte S, Leufkens HG, *et al*. Risk of cataract and glaucoma in patients with multiple sclerosis. Mult Scler. 2012;18(5):628-38.

Chapter 5:

The agreement between English primary care data and national death registration data in capturing vital status and the date of death

Chapter 5.1

The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations

Arlene M. Gallagher, Daniel Dedman, Shivani Padmanabhan, Hubert G. M. Leufkens, Frank de Vries

Pharmacoepidemiol Drug Saf. 2019;28(5):563-9

Abstract

Purpose: It is not clear whether all deaths are recorded in the Clinical Practice Research Datalink (CPRD) or how accurate a recorded date of death is. Individual level linkage with national data from the Office for National Statistics (ONS) and Hospital Episode Statistics (HES) in England offers the opportunity to compare death information across different data sources.

Methods: Age-standardised mortality rates (ASMRs) standardised to the European Standard Population (ESP) 2013 for CPRD were compared to figures published by the ONS and crude mortality rates were calculated for a sample population with individual linkage between CPRD, ONS and HES data. Agreement on the fact of death between CPRD and ONS was assessed and presented over time from 1998 to 2013.

Results: There were 33 997 patients with a record of death in the ONS data; 33 389 (98.2%) of these were also identified in CPRD. Exact agreement on the death date between CPRD and the ONS was 69.7% across the whole study period, increasing from 53.4% in 1998 to 78.0% in 2013. By 2013, 98.8% of deaths were in agreement within +/-30 days.

Conclusions: For censoring follow-up and calculating mortality rates, CPRD data are likely to be sufficient, as a delay in death recording of up to one month is unlikely to impact results significantly. Where the exact date of death or the cause are important, it may be advisable to include the individually linked death registration data from the ONS.

Keywords: CPRD, data linkage, death, HES, mortality, ONS, pharmacoepidemiology, validation, UK

Introduction

Electronic Health Records (EHRs) are increasingly used in clinical research. The Clinical Practice Research Datalink (CPRD) is an ongoing primary care database of anonymised medical records from general practitioners (GPs) and has been the basis of 886 published papers in the last 5 years [1]. The CPRD GOLD database is based on GP records using the Vision software system and the quality of primary care data is variable since data are entered by GPs during routine consultations, and not for the purpose of research.

Death is a common outcome measure in EHR studies with multiple articles published each year using CPRD GOLD [2-7]. Clear information on the date of death is important for censoring in observational epidemiology, especially for analyses of mortality and for studies assessing events at the end of life [8]. It is not clear whether all deaths are recorded in CPRD GOLD or how accurate a recorded date of death is [9]. In England and Wales, when a patient dies it is the statutory duty of the doctor who had attended in the last illness to issue the death certificate [10]. However, GPs may not routinely receive information about their deceased patients for whom they did not complete the death certificate [11]. Primary care software systems allow the recording of death information via multiple methods and changes to the software over time, or when converting between systems may impact the completeness of records [12]. CPRD researchers have developed an algorithm to identify multiple potential records of death from across the primary care record and combine this information to identify the best estimate for the date of death; this has not been externally validated (see appendix S1 for details of the algorithm).

Individual level linkage with national data from the Office for National Statistics (ONS) and Hospital Episode Statistics (HES) in England offers the opportunity to compare death information across the sources. This external validation will help researchers using EHRs to make decisions over the choice of data source when death is an important study variable. The objective of this study was to compare death date information based on the CPRD GOLD algorithm to national death registration data from the ONS and hospital death records from the HES for England. We examined how the mortality rate for CPRD GOLD compared to the published figures from the ONS and used individually linked data to investigate how both the fact of death and date of death compared between data sources.

Methods

The data used for this study were sourced from three routinely collected linked data sources in England. CPRD GOLD primary care data comprise the anonymous longitudinal EHRs of over 14 million patients in the UK in a dynamic cohort where patients can join/leave a GP practice over the course of follow-up [13,14]. These data are based on the Vision software system and have been shown in numerous validation studies to be generally of high quality [15,16]. Data have been collected since 1987. The ONS death registration data contains the date and coded cause of death for the population of England and Wales from January 1998, based on the death certificate and is considered the gold-standard [17]. The national HES data contain details of all admissions to National Health Service (NHS) hospitals in England from April 1997, including information on deaths in hospital [18]. Linkage between CPRD GOLD data and the ONS and HES data sources is available at an individual-level for those patients registered at practices in England that have consented to data linkage. Linkage between data sets is undertaken using a deterministic linkage algorithm, based on a patient's exact NHS identification number, sex, date of birth, and residential postcode.

A sample of one million patients from CPRD GOLD was identified in January 2014 for a previous publication, based on those who had been registered at a contributing practice for at least one day before this date [14]. We used the same sample, restricting it further to those alive when linked ONS death registration data became available (1998; 85.7%), and to those in England (81.5%) who were eligible for linkage with both the ONS data and the HES data (76.8%) to create our study cohort (see Figure SA). Follow-up time was assessed based on the time since registering at the GP practice until the patient was deducted from the record (transferred out of the practice for any reason). Analyses were restricted to the overlapping coverage period of active follow-up in all three data sources since the influence of the denominator definition is important in studies of linked data [19].

5.1

In order to compare the overall mortality rate in CPRD GOLD to the ONS data, we calculated age-standardised mortality rates (ASMRs) standardised to the European Standard Population (ESP) 2013 and compared this to figures published by the ONS using the same methodology. Data from the ONS were available from 2001 onwards. The ESP is an artificial population structure which is used in the weighting of mortality data to produce comparable ASMRs [20]. The population with individual linkage between all three data sources was used to calculate crude mortality rates in each data source and compare them. Poisson regression was used to calculate rates per 1000 person-years overall and by calendar year from 1998 onwards. We assessed agreement on the fact of death between the CPRD GOLD and the ONS and compared the difference in dates between the two over time.

Results

Table 1 describes the random sample of patients with data from 1998 to 2014 (n=857,047) and the study cohort of 536 341 in England who were eligible for linkage with both the ONS death registration data and the HES data; there were no major differences between the two. In the study cohort, the average follow-up time since registration at the GP practice was 11.2 years (standard deviation (SD) 12.4), of which 6.8 years (SD 5.5) were during the linkage coverage period. There were slightly more females than males (51.5% vs 48.5%). Most registered with their GP in the early years of life with over half registering before their 30th birthday. Body Mass Index (BMI) data was available for 70.7% of adult patients (aged 18 years or more), where the mean was 25.9 (SD 5.2); BMI information was missing for 97.9% of children. Smoking information was captured for 85.2% of adult patients; 25.4% were non-smokers, 29.6% ex-smokers and 30.2% smokers; Smoking information was missing for 86.4% of children and 11.0% were recorded as non-smokers. All patients in the sub-cohort available for linkage were in England, spread across all regions.

ASMRs per 100 000 population, standardised to the European Standard Population 2013 are presented separately for males and females in CPRD GOLD alongside the published national rates from the ONS, which were available from 2001 onwards (Figure 1). In both data sources the mortality rate was higher for males than females. The mortality rate published by the ONS was higher than calculated for CPRD GOLD for both males and females.

For our study cohort, the crude mortality rate for 1998 to 2013 using the ONS data was 9.33 per 1000 person-years (95% confidence interval 9.23 - 9.43). For the same patients, this compared to 9.41 (9.31 - 9.51) for CPRD GOLD data and 4.27 (4.21 - 4.34) for HES data. Mortality rates by calendar time for all three sources are presented in Figure 2.

There were 33 997 patients with a record of death in the ONS data; 33 389 (98.2%) of these were also in CPRD GOLD (Figure 3). There were 608 patients with a death recorded in the ONS but not in CPRD GOLD; 498 (82%) had been transferred out of the practice for a reason other than due to death. There were 834 additional patients for whom the algorithm used by CPRD identified a date of death, however there was no record in the ONS; 802 (96.2%) had been transferred out due to death. There were 17 deaths documented in HES that were not recorded in either the ONS or CPRD GOLD data.

Exact agreement between the death date estimated by the CPRD GOLD algorithm and the date from the ONS was 69.7% across the whole study period, increasing from 53.4% in 1998 to 78.0% in 2013 (Figure 4). An earlier death date was identified in CPRD GOLD for less than 3% and this did not change over time. A later date was identified in CPRD GOLD for 27.7%; reducing from 44.9% in 1998 to 19.5% in 2013. By 2013, 98.8% of deaths were in agreement within +/-30 days.

Table 1: Characteristics of a random sample of individuals in CPRD GOLD, active (alive) since 1998 and the subcohort in England eligible for linkage

Total N, %, unless stated otherwise	**Random sample of patients**	**Subcohort of the random sample: Patients in England, eligible for ONS and HES linkage**
Total number of patients (N)	857,047	536,341
Time since registration at the GP practice		
Time in years, mean (SD)	11.7 (12.9)	11.2 (12.4)
Time in years, median (IQR)	7.1 (2.3-17.2)	6.7 (2.3-16.5)
Follow-up time during the linkage coverage period		
Follow-up time in years, mean (SD)		7.3 (5.7)
Follow-up time in years, median (IQR)		5.8 (2.0-13.2)
Sex		
Female	442,055 (51.6%)	276,384 (51.5%)
Male	414,992 (48.4%)	259,957 (48.5%)
Age at registration with the GP practice		
Age in years, mean (SD)	28.1 (21.7)	28.4 (21.7)
Age in years, median (IQR)	27.0 (9.0-41.0)	27.0 (10.0-41.0)
Age <18	271,091 (31.6%)	165,438 (30.8%)
Age 18-29	213,520 (24.9%)	134,027 (25.0%)
Age 30-39	146,978 (17.1%)	94,554 (17.6%)
Age 40-49	84,790 (9.9%)	53,868 (10.0%)
Age 50-64	79,505 (9.3%)	50,021 (9.3%)
Age 65-74	31,534 (3.7%)	19,712 (3.7%)
Age 75+	29,629 (3.5%)	18,721 (3.5%)

BMI: body mass index; CPRD GOLD: Clinical Practice Research Datalink primary care database; GP: general practitioner; HES: Hospital Episode Statistics; IQR: Interquartile range; ONS: Office for National Statistics; SD: standard deviation.

Table 1 (cont.): Characteristics of a random sample of individuals in CPRD GOLD, active (alive) since 1998 and the subcohort in England eligible for linkage

Total N, %, unless stated otherwise	Random sample of patients	Subcohort of the random sample: Patients in England, eligible for ONS and HES linkage
Body Mass Index (BMI), aged 18+		
Mean BMI (SD)	25.9 (5.3)	25.9 (5.2)
Median BMI (IQR)	25.1 (22.3-28.6)	25.0 (22.3-28.5)
BMI unknown or out of range (<10, 70+)	211,792 (30.4%)	127,437 (29.3%)
Practice Region		
North East	13,158 (1.5%)	10,147 (1.9%)
North West	90,440 (10.6%)	73,149 (13.6%)
Yorkshire & The Humber	29,628 (3.5%)	19,430 (3.6%)
East Midlands	30,160 (3.5%)	16,758 (3.1%)
West Midlands	68,561 (8.0%)	55,029 (10.3%)
East of England	80,799 (9.4%)	61,526 (11.5%)
South West	71,233 (8.3%)	64,435 (12.0%)
South Central	98,990 (11.6%)	69,935 (13.0%)
London	126,807 (14.8%)	99,927 (18.6%)
South East Coast	88,714 (10.4%)	66,005 (12.3%)
Northern Ireland	20,677 (2.4%)	0 (0.0%)
Scotland	74,961 (8.7%)	0 (0.0%)
Wales	62,919 (7.3%)	0 (0.0%)

BMI: body mass index; CPRD GOLD: Clinical Practice Research Datalink primary care database; GP: general practitioner; HES: Hospital Episode Statistics; IQR: Interquartile range; ONS: Office for National Statistics; SD: standard deviation.

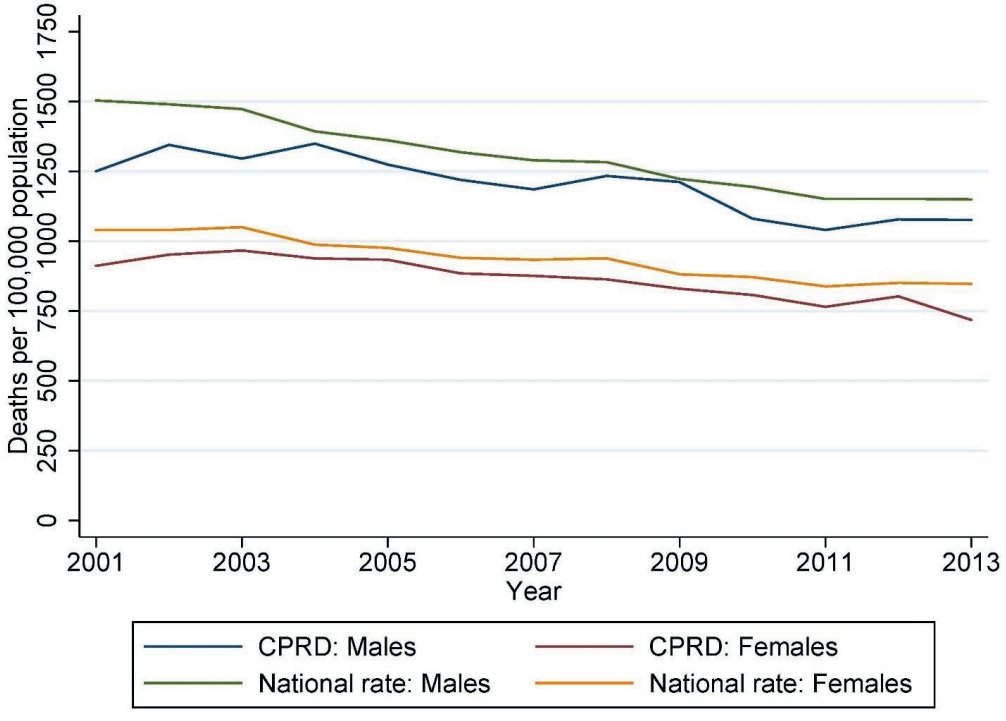


Figure 1: Age-standardised mortality rates (ASMRs), 2001 to 2013 for England

CPRD: Clinical Practice Research Datalink.

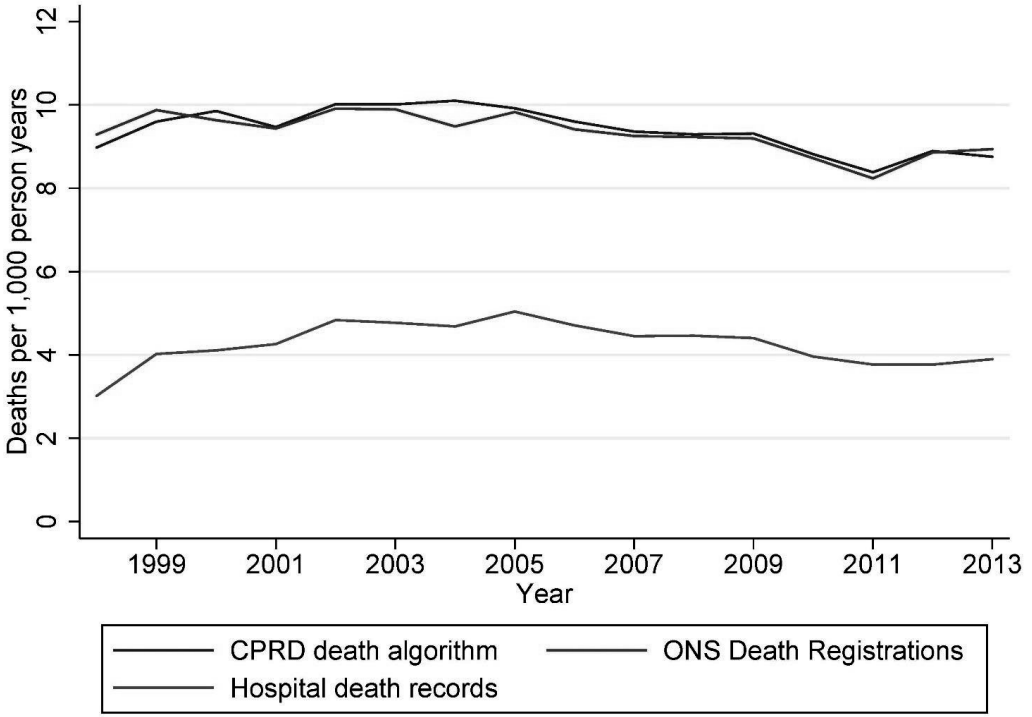


Figure 2: Mortality rates, 1998 to 2013 for the same patients in England, as recorded in three data sources

CPRD: Clinical Practice Research Datalink; ONS, Office for National Statistics.

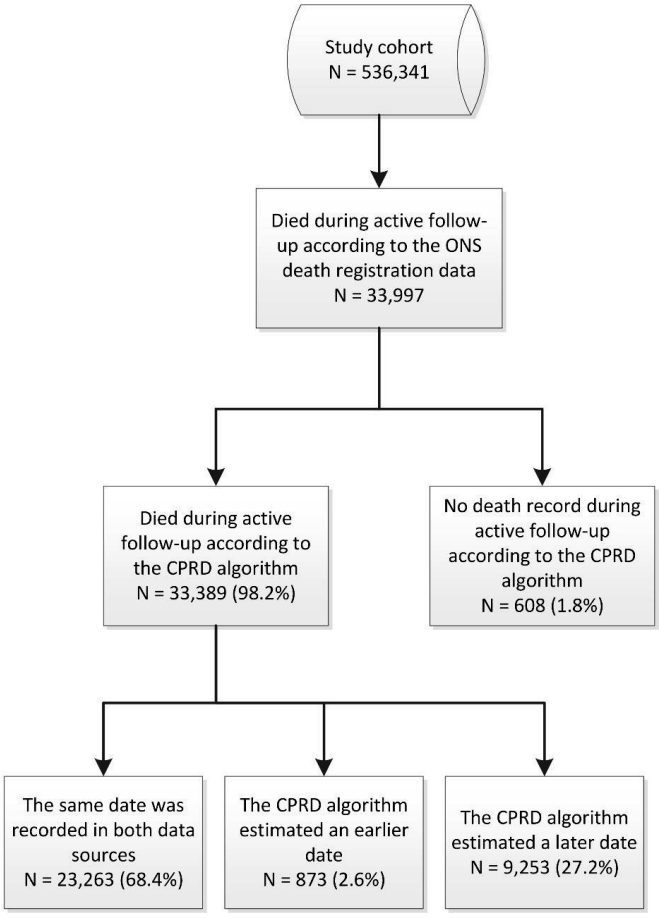


Figure 3: Concordance between the Clinical Practice Research Datalink (CPRD) algorithm and the gold-standard Office for National Statistics (ONS) data using individually linked data

5.1

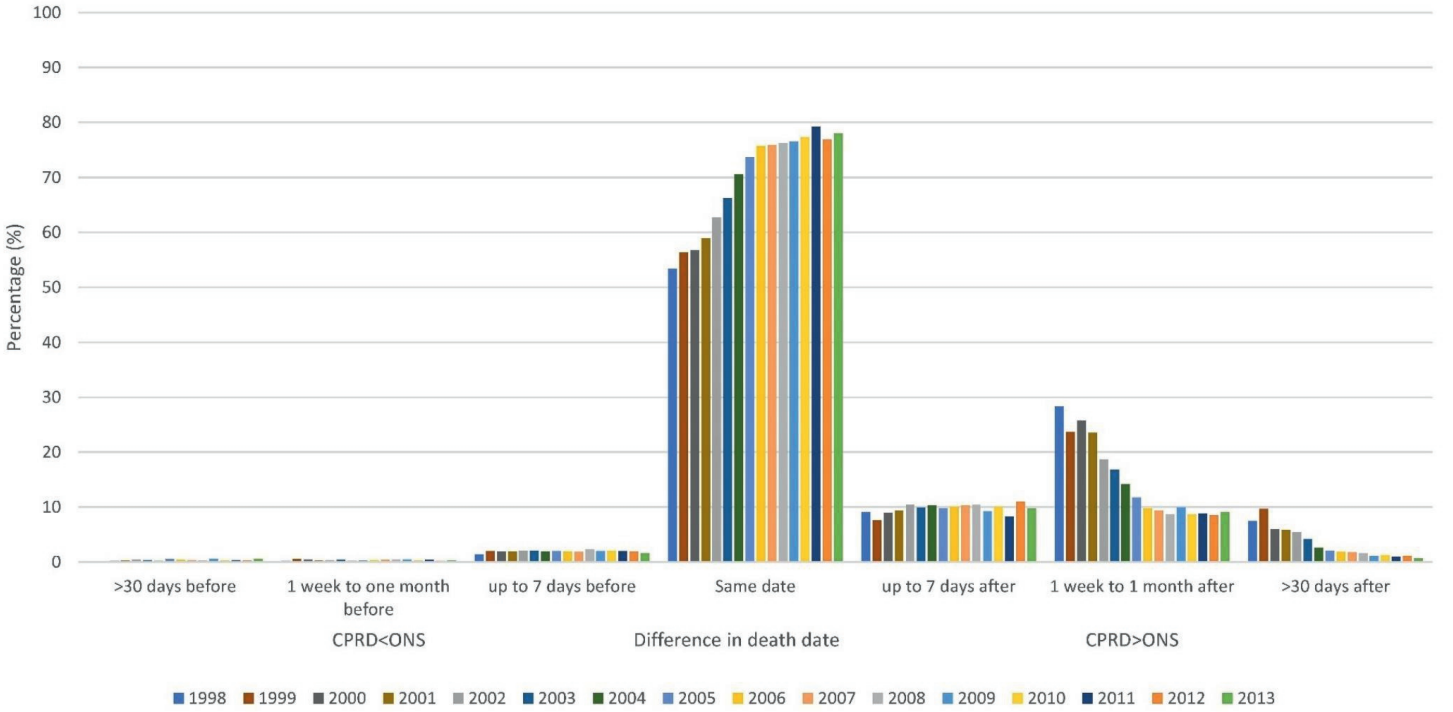


Figure 4: Difference between the CPRD death date and the Office for National Statistics (ONS) death date

CPRD: Clinical Practice Research Datalink; ONS: Office for National Statistics.

Discussion

This study showed that 98.2% of deaths in the ONS data are recorded in the CPRD GOLD primary care data. With such a high overall agreement rate, for studies where the fact of death is important, it would be acceptable to use death information from CPRD GOLD. Agreement on the exact date of death increased over time to 78.0% in 2013. Discrepancies were most commonly when the death date recorded at the GP practice was later than the official date on the death certificate, however both the amount of disagreement and the size of the delay has been reducing over time.

The objective of this study was to compare death date information based on the CPRD GOLD algorithm to national information from the ONS and hospital records, since GPs may not routinely receive information about their deceased patients and there is more than one way to record a death using the primary care software system. We were able to compare the information at the GP practice to the death certificate data by comparing individual records in both data sources.

The CPRD GOLD algorithm has not been externally validated and questions have been posed as to whether all deaths are recorded at the practice. The high level of agreement on the fact of death found in this study is encouraging and suggests that the majority of deaths are recorded in CPRD GOLD. Of the 1.8% that were not recorded in CPRD GOLD, a large proportion (82%) had been transferred out of the practice for another reason (not due to death), suggesting their record had been closed before the GP had been informed of their death. The HES data were limited to patients admitted to hospital, where as many as half of all deaths may occur [21]. For 2015/2016, 46.4% of deaths occurred in hospital in England [22] and we found a similar rate in this study. For this reason, we would not recommend using HES as the sole source of data for identifying death as an outcome.

This study found that even in recent years, one fifth of patients have a date of death recorded at the GP practice that is later than the date recorded in the ONS, which is based on the official death certificate. There are several reasons why there may be delayed recording of death information at the GP practice. Firstly, the practice may not be informed within a reasonable timeframe. For expected deaths, the registered GP is likely to be informed as soon as is practicable however, it is not essential for them to attend to verify a death [23] and the procedure for ensuring the registered GP is notified is not clear. They may be informed by the attending GP, the hospital, relatives, or information may first arrive from a central request to deduct the patient's record. Secondly, there may be limited incentive at the practice to close a patient's record immediately since their income relates to the list size [24,25]. Thirdly, there may be little reason to ensure the exact date of death is recorded since further clinical care is no longer required and this type of administrative task is unlikely to be

high priority for the practice. The lag time may be affected by the cause of death, which we did not assess in this study. For some conditions, patients are highly likely to be in hospital at the time of death, as was found for heart failure [26] and certain types of death, such as suicide, may be recorded differently [27]. Further studies could investigate the differences in lag time by cause of death. Delayed recording at the GP practice may contribute to an immeasurable time bias where CPRD GOLD is used for a study without linking to the ONS death registration data [28]. This type of bias is likely to have more impact on case-control study designs that use mortality as an outcome. In this study, we found the median delay to be 16 days. For studies where it is important to be exact on the date of death, it may be advisable to use the linked ONS data in order to minimise including immeasurable person time, however the impact is likely to be less in more recent years as both the proportion of patients and the median delay have reduced over time. This is in line with the improvement in the completeness of recording of many types of record at the GP practice in recent years [14].

This study used the same random sample of patients from a previous publication, which provided a detailed overview of the data available in CPRD GOLD [14]. The sample was limited to patients eligible for linkage with other data sources, and to those with follow-up since 1998. This immediately reduced the original sample of 1 million patients to 53% in a similar way to previously published studies [29]. Researchers need to balance in the loss in patient numbers alongside the gain in information from the linked data. Several CPRD GOLD studies have used the linked ONS data in order to provide additional information on the cause of death which would otherwise be less complete in primary care [30,31,32]. However, some studies have specifically chosen not to include linked ONS data even when death was an important study outcome, in order to maximise patient numbers and increase external validity by including all four countries of the UK [7]. In this study, no major differences were found between the original random sample of patients and the sub-cohort eligible for linkage with both the ONS and the HES data, who were based in England only. In addition, previous studies have shown no difference between mortality rates in a broad CPRD GOLD cohort compared to a subset eligible for linkage [33].

The ONS death registration data are based on official information from the Medical Certificate of Cause of Death (MCCD) and is seen as the gold-standard, as all deaths in the UK should be registered. Comparing this to the information available in CPRD GOLD provided the opportunity to see at the individual level how accurate the data at the GP practice is. However, there are some limitations to the ONS data. The ONS data include only deaths that occur in England and Wales, and not those that occur in other parts of the UK or abroad. Our patient sample may have included patients who died elsewhere, which may explain the small number of deaths which did not

appear in the ONS records. Delays in the registration of deaths can have an impact on the ONS mortality statistics [34]. Firstly, there can be delays in the issuing of the MCCD. In order to complete the MCCD, the certifying doctor must have seen the deceased in the previous two weeks of life. Delays can occur when there is a need for a post-mortem and/or inquest or when the death has been referred to a coroner, such as deaths with unknown cause, violent, unnatural, or suspicious deaths, those as a result of an accident, neglect, or suicide, or those which were employment related, occurred during an operation or when in custody [35]. Around 10,000 deaths a year in England and Wales are not registered for six months [36]. A very small number (0.3%) of deaths remain legally uncertified [35]. In addition, there is also a lag time between the issue of the MCCD and the record appearing in the ONS data. There is a 5-day legal limit on registering a death in England and Wales, which is usually done by the next of kin or a family member at a register office. However there has been an increase in the number missing this deadline in recent years due to difficulties in getting an appointment [37]. Future studies could investigate the difference between the date of death and the date of registration of the death.

To our knowledge, this is the first study to directly compare the recording of the date of death in primary records with the gold-standard from the ONS. Previous papers have relied upon comparing overall mortality rates [38], probabilistic linkage [39] or contacting GP practices for more information [40].

This study showed that the CPRD GOLD algorithm provides comparable results to using the ONS death registration data for the date of death. For censoring follow-up and calculating mortality rates, the primary care data are likely to be sufficient as a delay in death recording of up to one month is unlikely to impact results significantly. Where the exact date of death or the cause are important, it may be advisable to use the individually linked national ONS death registration data. Researchers should take into consideration the impact on sample size and power of a study when deciding whether to use the linked data as this can restricted the cohort size substantially.

Key points

- The majority of deaths are recorded in UK primary care data.
- The mortality rate calculated for CPRD is comparable with the ONS.
- Where the exact date of death or the cause is important in a CPRD study, it may be advisable to include the individually linked national ONS death registration data.
- Researchers should take into consideration the impact on sample size and power of a study when deciding whether to use the linked data as this can restrict the cohort size substantially.

References

1. Kontopantelis E, Stevens RJ, Helms PJ, Edwards D, Doran T, Ashcroft DM. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. BMJ Open. 2018;8(2):e020738.
2. Alatorre CI, Hoogwerf BJ, Deeg MA, Nelson DR, Hunter TM, Ng WT, Rekhter MD. Factors associated with stroke, myocardial infarction, ischemic heart disease, unstable angina, or mortality in patients from real world clinical practice with newly diagnosed type 2 diabetes and early glycemic control. Curr Med Res Opin. 2018;34(2):337-43.
3. Lane DA, Skjøth F, Lip GYH, Larsen TB, Kotecha D. Temporal Trends in Incidence, Prevalence, and Mortality of Atrial Fibrillation in Primary Care. J Am Heart Assoc. 2017;6(5):e005155.
4. Parisi R, Webb RT, Carr MJ, Moriarty KJ, Kleyn CE, Griffiths CEM, Ashcroft DM. Alcohol-Related Mortality in Patients With Psoriasis: A Population-Based Cohort Study. JAMA Dermatol. 2017;153(12):1256-62.
5. Stewart D, Han L, Doran T, McCambridge J. Alcohol consumption and all-cause mortality: an analysis of general practice database records for patients with long-term conditions. J Epidemiol Community Health. 2017;71(8):729-735.
6. Strongman H, Korhonen P, Williams R, Bahmanyar S, Hoti F, Christopher S, Majak M, Kool-Houweling L, Linder M, Dolin P, Heintjes EM. Pioglitazone and risk of mortality in patients with type 2 diabetes: results from a European multidatabase cohort study. BMJ Open Diabetes Res Care. 2017;5(1):e000364.
7. Zghebi SS, Steinke DT, Carr MJ, Rutter MK, Emsley RA, Ashcroft DM. Examining trends in type 2 diabetes incidence, prevalence and mortality in the UK between 2004 and 2014. Diabetes Obes Metab. 2017;19(11):1537-45.
8. Charlton J, Ravindrarajah R, Hamada S, Jackson SH, Gulliford MC. Trajectory of Total Cholesterol in the Last Years of Life Over Age 80 Years: Cohort Study of 99,758 Participants. J Gerontol A Biol Sci Med Sci. 2017;73(8):1083-9.
9. Harshfield A, Abel G, Barclay S, Payne R. Do GPs accurately record date of death? A UK observational analysis. BMJ Support Palliat Care. 2018;0:1-8.
10. Death Certification Advisory Group. Guidance for doctors completing Medical Certificates of Cause of Death in England and Wales. Office for National Statistics' Death Certification Advisory Group, Revised July 2010. Available at: https://www.gro.gov.uk/Images/medcert_July_2010.pdf [Accessed 18 July 2017]
11. Wagstaff R, Berlin A, Stacy R, Spencer J, Bhopal RA. Information about patients' deaths: general practitioners' current practice and views on receiving a death register. Br J Gen Pract. 1994;44(384):315-6.
12. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. Pharmacoepidemiol Drug Saf. 2009;18(1):76-83.
13. Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK primary care data resource. Ther Adv Drug Saf. 2012;3(2):89-99.
14. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36.
15. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69(1):4-14.

16. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract. 2010;60(572):e128-e136.
17. NHS Digital. Mortality data from the Office for National Statistics. Available at: http://content.digital.nhs.uk/onsmortality [Accessed 23 December 2016].
18. NHS Digital. Hospital Episode Statistics. Available at: http://content.digital.nhs.uk/hes [Accessed 23 December 2016].
19. Gallagher AM, Williams T, Leufkens HGM, de Vries F. The impact of the choice of data source in record linkage studies estimating mortality in venous thromboembolism. PLOS ONE. 2016;11(2): e0148349.
20. Eurostat. Revision of the European Standard Population: Report of Eurostat's task force, 2013 Available at: http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-RA-13-028 [Accessed 15 February 2018]
21. Public Health England. Recent trends in place of death in England. Available at: http://www.endoflifecare-intelligence.org.uk/view?rid=893 [Accessed 7 March 2017]
22. National End of Life Care Intelligence Network. Number and proportion of deaths by place of occurrence. Available at: http://www.endoflifecare-intelligence.org.uk/data_sources/place_of_death [Accessed 20 February 2018]
23. BMA. Confirmation and certification of death. Available at: https://www.bma.org.uk/advice/employment/gp-practices/service-provision/confirmation-and-certification-of-death [Accessed 20 February 2018]
24. NHS Digital. NHS Payments to General Practice: England, 2015/16. 21 September 2016. Available at: http://digital.nhs.uk/media/29211/NHS-Payments-to-General-Practice-England-2015-16-Report/Any/nhspaymentsgp-15-16-rep [Accessed 20 February 2018]
25. NHS Digital. General Practice trends in the UK to 2016. 19 September 2017. Available at: http://content.digital.nhs.uk/media/25201/General-Practice-Trends-in-the-UK-to-2016/pdf/General_Practice_Trends_in_the_UK_to_2016.pdf [Accessed 20 February 2018]
26. Hollingworth W, Biswas M, Maishman RL, Dayer MJ, McDonagh T, Purdy S, Reeves BC, Rogers CA, Williams R, Pufulete M. The healthcare costs of heart failure during the last five years of life: A retrospective cohort study. Int J Cardiol. 2016;224:132-8.
27. Thomas KH, Davies N, Metcalfe C, Windmeijer F, Martin RM, Gunnell D. Validation of suicide and self-harm records in the Clinical Practice Research Datalink. Br J Clin Pharmacol. 2013;76(1):145-57.
28. Suissa S. Immeasurable time bias in observational studies of drug effects on mortality. Am J Epidemiol. 2008;168(3):329-35.
29. Spaetgens B, de Vries F, Driessen JHM, Leufkens HG, Souverein PC, Boonen A, van der Meer JWM, Joosten LAB. Risk of infections in patients with gout: a population-based cohort study. Sci Rep. 2017;7(1):1429.
30. Wing K, Bhaskaran K, Smeeth L, van Staa TP, Klungel OH, Reynolds RF, Douglas I. Optimising case detection within UK electronic health records: use of multiple linked databases for detecting liver injury. BMJ Open. 2016;6(9):e012102.
31. Glover G, Williams R, Heslop P, Oyinlola J, Grey J. Mortality in people with intellectual disabilities in England. J Intellect Disabil Res. 2017;61(1):62-74.
32. Ratib S, Fleming KM, Crooks CJ, Walker AJ, West J. Causes of death in people with liver cirrhosis in England compared with the general population: a population-based cohort study. Am J Gastroenterol. 2015;110(8):1149-58.

33. Springate DA, Ashcroft DM, Kontopantelis E, Doran T, Ryan R, Reeves D. Can analyses of electronic patient records be independently and externally validated? Study 2--the effect of β-adrenoceptor blocker therapy on cancer survival: retrospective cohort study. BMJ Open. 2015;5(4):e007299.
34. Devis T and Rooney C. The time taken to register a death. Population Trends. 1997 Summer;(88):48-55.
35. Office for National Statistics. User guide to mortality statistics, July 2017. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017 [Accessed 15 February 2018]
36. Hawkes N, BMJ. 2014;349:g4305.
37. BBC. Death-registering delays 'rise by 70%'. Available at: http://www.bbc.co.uk/news/uk-england-41080306 [Accessed 9 February 2018]
38. Ohlmeier C, Langner I, Garbe E, Riedel O. Validating mortality in the German Pharmacoepidemiological Research Database (GePaRD) against a mortality registry. Pharmacoepidemiol Drug Saf. 2016;25(7):778-84.
39. Sieswerda E, Font-Gonzalez A, Dijkgraaf MG, Geskus RB, Heinen RC, van der Pal HJ, van Leeuwen FE, Caron HN, Kremer LC, Reitsma JB. Studying Hospitalizations and Mortality in the Netherlands: Feasible and Valid Using Two-Step Medical Record Linkage with Nationwide Registers. PLoS One. 2015;10(7):e0132444;1-13.
40. Hall GC. Validation of death and suicide recording on the THIN UK primary care database. Pharmacoepidemiol Drug Saf. 2009;18(2):120-31.

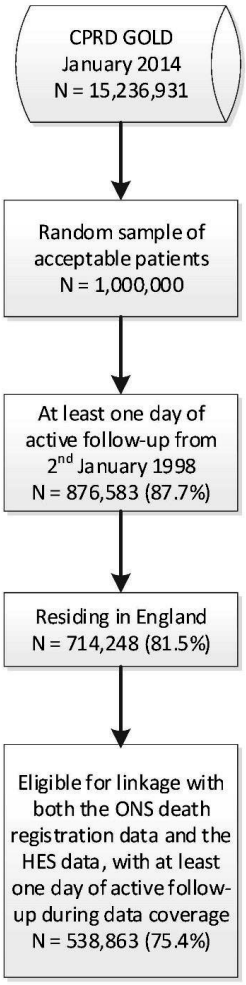


Figure SA: Cohort generation

CPRD GOLD: Clinical Practice Research Datalink primary care database; HES: Hospital Episode Statistics; ONS: Office for National Statistics

Appendix S1

Recording of the date of death in CPRD GOLD: the CPRD death algorithm

There is no single way to record the date of death using the Vision practice management system. The CPRD team has elaborated the following algorithm which aims to identify all deaths in the CPRD population.

Recording a death in the Vision practice management system
The following advice is given to practices about the recording of information relating to a death. When a practice is given notice that a patient has died, a Read code should be entered in the patient's medical history indicating that the patient has died, alongside the date of death, and the cause of death. Structured data areas are also available within Vision to enter specific information from the death certificate on death administration, or cause of death. At some point, the practice will close the patient's record by updating the patient's registration status to 'Transferred Out', using the reason 'Death'. There is often a delay between a patient's death and their record being closed.

Identifying candidate dates of death in CPRD GOLD
Since records relating to death can be logged in the clinical record, structured data areas, and via the patient registration status, the CPRD team developed an algorithm which aims to identify multiple potential death records and combine this information to identify the best estimate for the date of death.

For each patient, candidate death dates are identified from each of the following three sources within CPRD GOLD data.

1. Patient registration status marked as 'Transferred Out'

When a patient's record is closed, their registration status is updated which can include a transfer out date (tod) and a reason (toreason). Where the reason is specified as 'Death', the associated date provides an estimate for the date of death. These fields can be found in the Patient file.

5.1

2. Records in the death administration structured data area

Details of a death can be recorded using a structured data area for death administration (entity type 148). One of the fields is the date of death (data 1). This can be found in the Additional file. The associated date that this information was recorded (eventdate) can be found in the Clinical file. Both of these provide an estimate for the date of death.

3. Read codes indicating death in the patient's medical history

Read codes relating to death, suicide, postnatal death or neonatal death can be included in a patient's medical history. Details of the codes used by CPRD are listed in at the end of this appendix (Tables S1-S4). The associated date (eventdate) provides an estimate for the date of death. These fields can be found in the Clinical file.

For each of the four types of Read code there are different rules as follows:

- **Death** Read codes indicate that a patient has died. The date of the event (eventdate) is included in the algorithm.
- **Suicide** Read codes may refer to a completed suicide but may also refer to a suicide attempt, or to suicide of a family member. The date of the event (eventdate) is included in the algorithm so long as the patient also has a registration status marked as 'Transferred Out', where the reason is specified as 'Death' within the same or the next calendar year of the suicide record.
- **Neonatal death** Read codes may be recorded in the parent's, sibling's or the baby's record. The date of the event (eventdate) is only included in the algorithm if the patient's age at the event is under 2 years.
- **Postnatal death** Read codes may be recorded in the mother's or the baby's record. The date of the event (eventdate) is only included in the algorithm if the patient's age at the event is greater than 12 years.

The following events are ignored:

- Events dated prior to 1/1/1987
- Events which occurred prior to the patient's registration date at the practice
- Events which occurred after the last collection date from the practice
- Events with Read codes indicating death in the patient's medical history dated more than 95 days before the transfer out date, for patients with a registration status marked as 'Transferred Out'
- Events with Read codes indicating death in the patient's medical history for patients without either a registration status marked as 'Transferred Out', with the reason specified as 'Death' or a record in the death administration structured data area

Of the candidate death events identified from the three sources, the derived death date for the patient is initially set as the earliest event date from amongst the following:

- the transfer out date for patients with a registration status marked as 'Transferred Out', where the reason is specified as 'Death'
- the earliest of the date of death or date of recording of information in the death administration structured data area
- the earliest event date from Read codes indicating death in the patient's medical history.

The resulting date is added to the CPRD GOLD Patient file (deathdate).

Read codes by death category; September 2017 version

Table S1: Read codes used to identify neonatal death

Read code	Read code description
13M2.00	Death of infant
13M3.00	Sudden infant death
13MD.00	Death of child
22J5.00	O/E - dead - cot death
22J7.00	Postneonatal death
687N900	Sickle cell not screened for or screen incomplete, baby died
Q48y600	Early neonatal death
Q48y700	Late neonatal death
Q4z..11	Infant death
Q4z..12	Neonatal death
Q4z..13	Newborn death
Q4z..14	Perinatal death
R210.00	[D]Sudden infant death syndrome
R210000	[D]Cot death
R210100	[D]Crib death
R210200	[D]Nonspecific sudden infant death
R210z00	[D]Sudden infant death syndrome NOS
RyuC000	[X]Sudden infant death syndrome

Table S2: Read codes used to identify postnatal death

Read code	Read code description
13M8.00	Death of mother
L39A.00	Death obst cse occur more 42 day less than one yr aft deliv
L39B.00	Death from sequelae of direct obstetric causes
L39X.00	Obstetric death of unspecified cause
Lyu7500	[X]Obstetric death of unspecified cause
Q016.00	Fetus or neonate affected by maternal death
Q016.11	Fetus affected by maternal death

Table S3: Read codes used to identify death

Read code	Read code description
22J..00	O/E - dead
22J..11	O/E - dead - condition fatal
22J..12	Death
22J..13	Died
22J..14	Patient died
22J1.00	O/E - dead - unexpected
22J2.00	O/E - dead - expected
22J3.00	O/E - dead - unattended death
22J4.00	O/E - dead - sudden death
22J6.00	O/E - dead - suspicious death
22J8.00	Death - cause unknown
22JZ.00	O/E - dead NOS
38Qd.00	Gold Standards Framework After Death Analysis Audit Tool
4K9..00	Post mortem exam.
4K91.00	Post mortem exam. requested
4K92.00	Coroner's post mortem exam.
4K93.00	Home office post mortem
4K94.00	Post mortem exam. done
4K95.00	Post mortem result
4K96.00	Post mortem request
4K9Z.00	Post mortem exam. NOS
56C1.00	Post-mortem radiology normal
7L1M000	Preoperative anaesthetic death
8HG..00	Died in hospital
8HG..11	Death in hospital
9134.00	Registration ghost - deceased
9134.11	Registration ghost - dead
9134.12	Registration ghost - died
9234.00	FP22-death
94...00	Death administration
94...11	Administration after pat. died
941..00	Death certificate form Med A
941..11	Certificate - death
9411.00	Death cert. Med A due
9412.00	Death cert. Med A signed
9413.00	Med A given to family
9414.00	Med A not signed-coroner case
941Z.00	Death cert. Med A NOS
943..00	Report for Coroner
9431.00	Coroner report - requested
9432.00	Coroner report - sent off
9433.00	Coroner report - paid for
943Z.00	Report for Coroner NOS
944..00	Coroner's post-mortem report
944..11	Post-mortem report-police surg
9441.00	Coroner's PM report awaited
9442.00	Coroner's PM report requested
9443.00	Coroner's PM report received
944Z.00	Coroner's PM report NOS
945..00	Hospital death discharge notif
9451.00	Death notif. from hospital
9452.00	Await hosp death disch letter
9453.00	Receiv hosp death disch letter
9454.00	Ask for hosp death disch lett.
945Z.00	Hospital death disch. NOS
946..00	Death notif.- non.hosp source
947..00	Cause of death clarif. SD17/18
947..11	SD17 - cause of death clarif
947..12	SD18 - cause of death clarif
9471.00	SD17/18 received-death clarif.
9472.00	SD17/18 completed
9473.00	SD17/18-no details, returned
947Z.00	SD17/18 cause of death NOS
948..00	Cremation certification
948..11	Stat B,C and F cremation certs
9481.00	Patient for cremation
9482.00	Crem. form part B completed
9483.00	Crem. form part C arranged
9484.00	Crem. form part C completed
9485.00	Cremation form 4 completed
9486.00	Cremation form 5 completed
948Z.00	Cremation certification NOS
949..00	Patient died - to record place
949..11	Dead - place patient died
949..12	Deceased - place patient died
949..13	Died - place patient died
949..14	Place of death
9491.00	Patient died at home
9492.00	Patient died in part 3 accom.
9493.00	Patient died in nursing home
9494.00	Patient died in resid.inst.NOS
9495.00	Patient died in hospital
9496.00	Patient died in street
9497.00	Patient died in publ.place NOS
9498.00	Dead on arrival at hospital
9499.00	Found dead at accident site
949A.00	Patient died in hospice
949B.00	Patient died in community hospital
949C.00	Patient died in GP surgery
949D.00	Patient died in care home
949E.00	Patient died in usual place of residence
949F.00	Died in ambulance
949G.00	Patient died in hospice community lodge
949Z.00	Patient died in place NOS
94A..00	Unexpected death-Coroner told
94A..11	Referral to coroner
94B..00	Cause of death
94B..11	Condition fatal-cause of death
94C..00	Post mortem report

Table S3 (cont.): Read codes used to identify death

Read code	Read code description
94C0.00	Post mortem report received
94C1.00	Police surgeon postmortem report
94D..00	Hospital notified of death
94E..00	Date of death
94F..00	Unexpected death
94G..00	Verification of expected death
94Z..00	Death administration NOS
9m06.00	Excluded from diabetic retinopathy screening as deceased
G575100	Sudden cardiac death, so described
R2...00	[D]Cause of morbidity and mortality unsure and ill-defined
R2...12	[D]Mortality, cause unsure
R21..00	[D]Sudden death, cause unknown
R211.00	[D]Instantaneous death
R212.00	[D]Death less than 24 hours from onset of illness
R212000	[D]Death, not instantaneous cause unknown
R212100	[D]Died, with no sign of disease
R212z00	[D]Death less than 24 hours from onset of illness NOS
R213.00	[D]Unattended death
R213000	[D]Found after death, unknown cause of death
R213100	[D]Found dead
R213z00	[D]Unattended death NOS
R21z.00	[D]Sudden death, cause unknown NOS
R2y..00	[D]Cause of morbidity/mortality ill-defined or unsure OS
R2yz.00	[D]Other and unknown causes of morbidity or mortality NOS
R2z..00	[D]Cause of morbidity/mortality ill-defined and unsure NOS
RyuC.00	[X]Ill-defined and unknown causes of mortality
RyuC100	[X]Other sudden death, cause unknown
T053.00	Killed by rolling stock
T053100	Killed by rolling stock - passenger
T053200	Killed by rolling stock - pedestrian
T053z00	Killed by rolling stock - unspecified person
T0y0.00	Found dead on railway right-of-way unspecified
T0y0200	Found dead on railway unspecified - pedestrian
T0y0300	Found dead on railway unspecified - cyclist
T0y0z00	Found dead on railway unspecified - unspecified person
TGyz400	Accidentally killed NOS
TL60.00	Homicidal cut of any part of body
TL61.00	Homicidal puncture of any part of body
TL62.00	Homicidal stab of any part of body
TL9..00	Homicide
TLxz000	Manslaughter, nonaccidental, NOS
TLxz200	Assassination successful NOS
TLxz400	Murder successful NOS
U....00	[X]External causes of morbidity and mortality
U3...12	[X]Homicide
U3...13	[X]Murder
ZV68011	[V]Issue of death certificate

Table S4: Read codes used to identify suicide

Read code	Read code description
TK...00	Suicide and selfinflicted injury
TK...11	Cause of overdose - deliberate
TK...13	Poisoning - self-inflicted
TK...14	Suicide and self harm
TK0..00	Suicide + selfinflicted poisoning by solid/liquid substances
TK00.00	Suicide + selfinflicted poisoning by analgesic/antipyretic
TK01.00	Suicide + selfinflicted poisoning by barbiturates
TK01000	Suicide and self inflicted injury by Amylobarbitone
TK01100	Suicide and self inflicted injury by Barbitone
TK01400	Suicide and self inflicted injury by Phenobarbitone
TK01z00	Suicide and self inflicted injury by barbiturates
TK02.00	Suicide + selfinflicted poisoning by oth sedatives/hypnotics
TK03.00	Suicide + selfinflicted poisoning tranquilliser/psychotropic
TK04.00	Suicide + selfinflicted poisoning by other drugs/medicines
TK05.00	Suicide + selfinflicted poisoning by drug or medicine NOS
TK06.00	Suicide + selfinflicted poisoning by agricultural chemical
TK07.00	Suicide + selfinflicted poisoning by corrosive/caustic subst
TK0z.00	Suicide + selfinflicted poisoning by solid/liquid subst NOS
TK1..00	Suicide + selfinflicted poisoning by gases in domestic use
TK10.00	Suicide + selfinflicted poisoning by gas via pipeline
TK11.00	Suicide + selfinflicted poisoning by liquified petrol gas
TK1y.00	Suicide and selfinflicted poisoning by other utility gas
TK1z.00	Suicide + selfinflicted poisoning by domestic gases NOS
TK2..00	Suicide + selfinflicted poisoning by other gases and vapours
TK20.00	Suicide + selfinflicted poisoning by motor veh exhaust gas
TK21.00	Suicide and selfinflicted poisoning by other carbon monoxide
TK2y.00	Suicide + selfinflicted poisoning by other gases and vapours
TK2z.00	Suicide + selfinflicted poisoning by gases and vapours NOS
TK3..00	Suicide + selfinflicted injury by hang/strangulate/suffocate
TK30.00	Suicide and selfinflicted injury by hanging
TK31.00	Suicide + selfinflicted injury by suffocation by plastic bag
TK3y.00	Suicide + selfinflicted inj oth mean hang/strangle/suffocate
TK3z.00	Suicide + selfinflicted inj by hang/strangle/suffocate NOS
TK4..00	Suicide and selfinflicted injury by drowning
TK5..00	Suicide and selfinflicted injury by firearms and explosives
TK51.00	Suicide and selfinflicted injury by shotgun
TK52.00	Suicide and selfinflicted injury by hunting rifle
TK53.00	Suicide and selfinflicted injury by military firearms
TK54.00	Suicide and selfinflicted injury by other firearm
TK5z.00	Suicide and selfinflicted injury by firearms/explosives NOS
TK6..00	Suicide and selfinflicted injury by cutting and stabbing
TK60.00	Suicide and selfinflicted injury by cutting
TK61.00	Suicide and selfinflicted injury by stabbing
TK6z.00	Suicide and selfinflicted injury by cutting and stabbing NOS
TK7..00	Suicide and selfinflicted injury by jumping from high place
TK70.00	Suicide+selfinflicted injury-jump from residential premises
TK71.00	Suicide+selfinflicted injury-jump from oth manmade structure
TK72.00	Suicide+selfinflicted injury-jump from natural sites
TK7z.00	Suicide+selfinflicted injury-jump from high place NOS
TKx..00	Suicide and selfinflicted injury by other means
TKx0.00	Suicide + selfinflicted injury-jump/lie before moving object
TKx0000	Suicide + selfinflicted injury-jumping before moving object

Table S4 (cont.): Read codes used to identify suicide

Read code	Read code description
TKx0z00	Suicide + selfinflicted inj-jump/lie before moving obj NOS
TKx1.00	Suicide and selfinflicted injury by burns or fire
TKx2.00	Suicide and selfinflicted injury by scald
TKx3.00	Suicide and selfinflicted injury by extremes of cold
TKx4.00	Suicide and selfinflicted injury by electrocution
TKx5.00	Suicide and selfinflicted injury by crashing motor vehicle
TKx6.00	Suicide and selfinflicted injury by crashing of aircraft
TKx7.00	Suicide and selfinflicted injury caustic subst, excl poison
TKxy.00	Suicide and selfinflicted injury by other specified means
TKxz.00	Suicide and selfinflicted injury by other means NOS
TKz..00	Suicide and selfinflicted injury NOS
U2...13	[X]Suicide

Chapter 6:

General Discussion

The overall thesis aim was to evaluate the utility of the CPRD GOLD database for studies with death as an outcome. Each chapter contributed to addressing each of the four aims, considering the generalisability, appropriateness, quality and accuracy of CPRD data for mortality studies. Chapter 2 reviewed the primary care data from CPRD as a data source epidemiological studies and discussed generalisability with the UK and Dutch populations, in order to researchers decide whether to use this data source for pharmacoepidemiology studies with death as an outcome. Chapter 3 presented examples of studies where death was an outcome, based on primary care data alone and when linked with other data sources. Chapter 4 outlined the implications in terms of the denominator and cohort size when using linked data and chapter 5 looked in detail at the accuracy of vital status and the date of death in the primary care record.

In this chapter I address each of the 4 objectives by reviewing the key findings and placing the results in the context of related published studies. In addition, I make recommendations of areas for future research and outline where CPRD provides a valuable contribution in analyses of death before drawing final conclusions.

Aim 1: To review the CPRD as a data source for pharmacoepidemiological studies of mortality

The first aim was to review the CPRD as a data source for pharmacoepidemiological studies of mortality. Chapter 2 examined the primary care data collected and the representativeness of the population. The publication of chapter 2.1 as part of the Data Resource Profile series in the International Journal of Epidemiology provides open access to a general description of the database, which was not previously available. Before this was published, the most up to date reference for CPRD data was Williams *et al*, 2012 [1]. This had focused on a history of the GPRD, redevelopment of the database in 2009, the new linkage offering and methodological advances in the use of the data. Chapter 2.1 specifically focused on the primary care data, in order to support researchers in epidemiology to understand the scope of the data resource and how to access and make best use of the data. Chapter 2.1 provided an overview of the content of the primary care dataset, described the strengths (including representativeness and data quality) and weaknesses (including missing data), and outlined access to this resource. Since publication, this paper has been cited more than 600 times [2].

Chapter 2.1 outlined the strengths of this data source. The healthcare system in the UK, with the GP as the gatekeeper to non-emergency secondary care, provides a strong setting for research, and the large population size, breadth of recording in the coded medical history, collection of lifestyle data and longitudinal follow-up for patients, supports the use of this data source for studies in pharmacoepidemiology. The data collected in the CPRD primary care database was found to be representative on age and sex to the national census data however when studying mortality, researchers also need to assess the external validity of this data source, the data quality and consider the appropriateness of the primary care setting.

External Validity

In order to assess the utility of the database for pharmacoepidemiological studies, it is important to understand the representativeness of the database in terms of external validity. At the time of starting this thesis, questions had been raised on how well the CPRD primary care database reflected the whole of the UK and indeed compared to other European countries. It has largely assumed that data in the CPRD is comparable to the population of other western countries but there have not been any studies directly comparing the populations to back this up. The CPRD research team had presented only conference abstracts on representativeness without a full published paper. One abstract focused on whether the whole database was representative of the UK [3] and two abstracts described how the subset of the database that could be linked to other data sources compared to the whole database [4, 5]. An increasing number of CPRD studies are being conducted by researchers

from outside the UK, particularly in the Netherlands. There had not been any studies outlining the differences or similarities between these populations and therefore chapter 2.2 aimed to compare both the age and sex distribution of the CPRD database with the total Dutch population. Chapter 2.2 presented figures showing that overall the CPRD population are similar to the Netherlands in terms of age and sex. This unique study was a first step in showing that results from observational studies using CPRD data are applicable to the Dutch population. Together the two papers in chapter 2 provided reassurance for all studies where age and sex are major confounding factors.

Although chapters 2.1 and 2.2 provided confidence in the database in terms of age and sex, there are other important variables that influence the applicability of study results from a database, such as the representativeness in terms of geography and practice list size. Presenting robust information in these areas is difficult due to the dynamic nature of the database where practices join, leave, split or merge over time, and patients move between practices over their lifetime. However, we know that there is variation in all-cause mortality by UK region [6]. The ONS concluded that the substantial variation seen between different local areas reflects underlying geographical differences in factors such as income deprivation, socio-economic status and health behaviour. Study findings of worse health outcomes in some regions have reflected higher mean levels of deprivation [7]. It is therefore important that measures of deprivation are available for research when using the CPRD for mortality studies, particularly given the UK population is both growing and ageing [8] which is likely to impact all-cause mortality over time.

Chapter 2.1 only briefly mentioned the data not captured. Some populations are likely to be missing from the EHR, examples include prisoners, the armed forces and their families, and individuals residing in some nursing and care homes. Although the numbers are likely to be small, this could impact all-cause mortality if these groups have a different mortality risk. Firstly, in the UK, prisoners consult their GP three times more frequently than a demographically equivalent population in the community [9, 10]. The mortality rate amongst the prison population differs markedly from the general population, particularly for suicide where peak rates of suicide in 1999 and 2016 of 1.4 deaths per 1,000 prisoners, compares with rates of 0.16 per 1,000 in the general male population [11]. The prison population is reported as 0.13% of the UK population [12]. Secondly, the Ministry of Defence (MoD) is responsible for providing primary care services for servicemen and women within the UK and at defence outposts overseas [13]. Overall, in 2018, the UK Regular Armed Forces were at a statistically significant lower risk of dying compared to the UK general population [14]. The British armed forces and their reserves make up over 190,000 personnel in the population [15]. Thirdly, some nursing and care homes use their own GP, others register all their patients with the same GP therefore it is not clear how

well these populations are captured in the EHR. Polypharmacy in nursing home residents is a concern, where the prevention and management of age-related chronic conditions have resulted in an increased number of medications taken by older adults [16]. As the number of medications taken increases, so does the risk for adverse events, including death [17, 18]. In people aged 75 and over in 2006–08, 12.1% of deaths were in nursing homes and 10.0% were in old people's homes [19]. Other populations that may be missing from the CPRD include the homeless, illegal immigrants, permanent residents of psychiatric institutions and those of any age under hospice care. Users of the CPRD need to consider the impact of missing populations in the context of their study, for example, studying end of life care in CPRD, where patients in care homes or hospices may be underestimated.

Data quality

In addition to external validity, it is important to understand the quality of the data and the extent of potential missing information on exposures, outcomes and key covariates. Analyses of the database in chapter 2.1 found that there is variability in the completeness of data across patients and across time. Missing data in the EHR may have an impact on all-cause mortality studies due to exposure or outcome misclassification, or residual confounding. The Quality and Outcomes Framework (QOF) has increased the recording of lifestyle factors in EHRs in the UK since 2004 and these are important confounders in assessing all-cause mortality [20, 21]. Individuals commonly have their height, weight, blood pressure, smoking status, and alcohol consumption recorded during the first year of registration with a UK GP, however the proportion falls in the years that follow [22]. In addition, the recording of lifestyle factors is more likely to be missing for those who have no cause to visit the GP, a form of healthy user bias. Researchers need to consider the potential for residual confounding due to incomplete covariate capture using the CPRD.

Chapter 2.1 discussed misclassification in terms of missing data. There is a potential for exposure misclassification in pharmacoepidemiology studies using databases with missing data on treatments. Drug exposure in the CPRD is defined by prescriptions. There is no link between the primary care data and dispensing at the pharmacy and there is also missing information on over the counter drug use. Such missing information can lead to patients being erroneously categorised as unexposed. For cohort studies, exposure misclassification is likely to be non-differential, where the degree of misclassification of exposure status among those with and those without the outcome is the same; this is most likely to bias estimates toward no association. There is also the potential for outcome misclassification using CPRD data since researchers rely upon the presence of a code or record to indicate the presence of the condition or fact of death, and vice versa. For this thesis the key concern was the validity of the death records, particularly the information on death dates. Chapters 3 to 5 looked at the likely impact on mortality studies in more detail.

Setting

For most mortality studies, the UK primary setting will offer an appropriate pool of patients since almost all consultations are routed through the GP. Patients presenting in primary care are likely to be a good representation of clinical care in the community. The system of uniform national healthcare, with the GP as the gatekeeper to non-emergency secondary care, together with the potential for linkage between primary care data, hospital information and the registry of deaths, creates a strong data source for research. This setting would suit studies of conditions managed in primary care such as asthma and may provide more generalisable results on the association between asthma severity and deaths than found when using a hospital-based population [23]. However, delays in the relay of information between sources makes it difficult to study short term mortality following hospital procedures or deaths associated with acute onset events such as strokes or pulmonary embolism, where the robustness of the data relies upon retrospective recording at the GP practice.

In conclusion, chapter 2 presented a broad overview of the CPRD as a data source and provided confidence that this data source is representative of the UK and Dutch populations in terms of age and sex. This provides us with confidence in using this data source for mortality studies, where age and sex are important covariates. Researchers should consider the generalisability for studies of exposures or outcomes likely to be captured in secondary care or those more prevalent in populations not covered.

Aim 2: To conduct case studies demonstrating the use of primary care data alone for studies with a mortality outcome and when using linked data

The second aim was to present studies of mortality using CPRD data. Since linked data were not yet available at the start of this PhD, chapter 3.1 presented a study assessing mortality using primary care data only. Previous research had found that most high-risk AF patients who were candidates for anticoagulation, were either not taking warfarin or had a subtherapeutic INR at the time of admission to hospital with a stroke. Reduced anticoagulation control in warfarin users had already been associated with an increased risk of stroke and mortality, however these studies had small cohorts, were limited to patients in Wales or compared warfarin users to patients untreated with warfarin [24-26]. This study aimed to present the risks of stroke and mortality in a larger cohort of AF patients using warfarin, based on their time within therapeutic range when including patients from across the UK countries. CPRD primary care data was a useful data source to answer this question since both the Atrial Fibrillation diagnosis and exposure to warfarin are likely to be captured at the GP practice; AF is generally managed in primary care, unless treatment fails to control the symptoms and more specialised management is needed, and GPs are responsible for the prescribing of warfarin for the prevention of stroke.

The strength of this study was the cohort size; previous studies had analysed less than 3,000 AF patients with INR readings, whereas the study population in chapter 3.1 included over 27,000. In addition, the primary care database provided records on lifestyle factors, other antithrombotic treatments, duration of AF and socioeconomic status which could be included to adjust for confounding in the analyses. An additional benefit of this study over previous research was the ability to include patients from all four countries of the UK. However, the generalisability of the population is questionable since not all patients would have had a confirmed diagnosis of AF. Previous research on AF in Sweden excluded patients without an electrocardiogram (ECG) to confirm the diagnosis [27], whereas this study relied upon on a code for AF in the primary care record based upon a validation study by Ruigomez *et al.* (2002) [28] which reported a high level of validity in the recording of AF by GPs. The results of the study could therefore be generalised to AF patients who present in primary care, but may not represent those with more severe AF or poorly managed anticoagulation, who are more likely to present in secondary care and would have been missing from the inclusion criteria if they did not survive to be managed in primary care. The results from chapter 3.1 may therefore have been more biased towards AF patients who present in primary care and therefore potentially less at risk of death. At the time of this study it was not yet possible to link to the HES or national death records. The fact of death and the date recorded in the GP record was assumed to be recorded accurately as most primary care CPRD

studies either with a mortality outcome or using death as a right-censoring point, had not included any concerns in the study limitations of their publications.

At the time of the design of the study in chapter 3.2, linked data were available. Incident AF patients were identified from diagnoses in the primary care record and 39% were eligible for linkage, based on the consenting English practices and the recording of the necessary identifiers in the patient record. A comparison was made between the incident AF population in CPRD and the subset eligible for linkage. No major differences were found in age, sex, body mass index, alcohol use or smoking status, and the history of comorbidities, $CHADS_2$ risk scores and previous treatment with cardiovascular drugs in the subset were similar to the wider cohort. This provided good evidence that the subset of incident AF patients eligible for linkage was representative of the wider incident AF population presenting in primary care. The only difference was in the length of follow-up; the median duration of follow-up in the overall cohort was 2.2 (range 0–6) vs 1.7 (0–5) years in the final linked AF cohort. This reflected the limited period for which the linked data were available (up to October 2009 for the HES and November 2009 for the ONS data versus November 2010 for primary care data). This study showed that researchers deciding whether to use linked data therefore must consider the impact on both the size of the cohort and the available follow-up time.

In chapter 3.2 the risks of cardiovascular events, bleeding and mortality were assessed within the linked cohort comparing periods of exposure to anticoagulants (current, recent, past or no use). The availability of linked data allowed all-cause mortality to be assessed using the official date of death from the ONS death registration data. However, no comparison was made between the records from ONS and the GP data. In addition, although linked HES data were used to assess hospitalisations and cardiovascular and bleeding events, these data were not used to confirm the diagnosis of AF or identify any additional patients presenting in secondary care, without a primary care diagnosis.

Chapter 3.3 also assessed time within therapeutic range, however for patients with VTE. The study assessed all-cause mortality and deaths due to VTE, compared to controls. This was the first study of VTE to combine primary care, hospital data and information from death certificates in the UK. The availability of linked data brought two advantages, i) classifying the cases and controls and ii) identifying the outcome events. The VTE diagnosis was based on coded diagnoses in either the GP record or in the hospital data. This enabled the study to include those with VTE records from secondary care even if they had no record at the GP practice. Previous studies had found that up to 20% of VTE cases may present in secondary care [29] and 8,918 cases were identified in the HES population linked to CPRD GOLD (data not shown). In addition, information from the free text area in the primary care data was available

at the time and was used to exclude patients with negating or questionable notes in relation to the VTE record (e.g., ‘‘?’’ or ‘‘ruled out’’); this cohort was therefore a robust reflection of the VTE population in the UK. In addition to adding further cases that would not have been identified in the primary care, the linked data were used to ensure control patients had no record of VTE in either primary or secondary care. Linked death registration data were available for 41% of patients and these made it possible to look at cause of death, specifically due to VTE, which may not have been well recorded in the primary care data. As with chapter 3.2, the linked death data were assumed the gold-standard and analyses of the data available in primary care for all-cause mortality were not presented in chapter 3.3. Our first opportunity to examine the rates from both sources for the same cohort was in chapter 3.4.

In the final case study, chapter 3.4 presented a study of patients treated with insulin or oral antidiabetic drugs. At the time of this study, there were concerns over the cardiovascular safety of the thiazolidinedione drugs based on meta-analyses of clinical trial data. It was likely that the patients using these treatments in actual clinical practice would be different from those in the clinical trial setting and previous observational studies had been limited by limited statistical power, short follow-up time or a lack of information in important confounders such as lifestyle factors. In this study, new users of pioglitazone, rosiglitazone, insulin, metformin or sulphonylureas were compared at baseline and analysed for cardiovascular and mortality outcomes using both primary care data and linked HES and ONS death registration data. In this study, mortality rates were calculated based on primary care data for the wider cohort and separately for the subset of patients eligible for linkage, using death certificate data. The incidence rate for deaths was similar for the subset with linked data, compared with the wider primary care cohort, however, the results were more precise for the larger cohort using GP data only.

One strength of this study over previous publications was the ability to adjust for confounding factors such as age, sex, calendar year, socioeconomic status, smoking status, use of alcohol, body mass index, medical history of cardiovascular events and previous prescribing of cardiovascular and anti-diabetic treatments in large inception cohorts by class of diabetes medication. However, a limitation was the reduction in cohort size for analyses of linked data. It was only possible to analyse the cause of death in the subset of patients linked to ONS death registration data. The substantial restriction on the number of patients had an impact when assessing the cause of death by ICD-10 chapter, where for some causes there were less than 5 individuals. Since the time of publication, the CPRD have introduced a policy that no cell should contain <5 events, therefore for future studies like this, some results would have to be supressed.

These studies showed that there are both opportunities and challenges arising from the availability of individually linked data for mortality studies using UK primary care EHRs. Linked HES data provide additional information on diagnoses from hospitals, enabling studies to include patients with records in secondary care which may not have been reflected in the EHR from the GP practice. This is particularly useful for diagnoses such as AF and VTE, for which patients may present in either primary or secondary care but confirmation of the diagnosis is usually made by specialists in a hospital setting. Linked ONS death registration data provide further details on the death record, such as the cause(s) and may identify deaths in the study cohort that have not been reflected in the EHR. However, linked data could not resolve all limitations when using primary care data. For all three of these linked studies (chapters 3.2-3.4), assumptions were made on the length of prescriptions as there were limited details in the primary care record and no linkage was available to more robust data on pharmacy dispensing. Similarly, missing data on over the counter treatments could not be corrected for with linked data as these data are not recorded in a database available for linkage. Over the counter aspirin use may have an impact on anticoagulation management for the AF and VTE patients. In addition to missing exposure data, specialist clinics, such as those managing anticoagulation or diabetes, may not be held at the GP practice and details from these may not be passed to the GP practice in good time, or at all. Chapters 3.1, 3.2 and 3.3 all examined anticoagulation via INR measurements. Time within therapeutic range could only be calculated for patients with at least three measurements of INR. It is possible that some patients were excluded where their measurements were taken outside of the GP practice. In addition, intermittent missing information on INR measurements limits the ability to calculate time in therapeutic range accurately, particularly if measurements are far apart over the study period. There are methodological issues when attempting to estimate the proportion of time a patient is within or outside the INR target range. The method of linear interpolation [30] has been shown to be the most valid and simplest technique compared with other methods [31]. This assumes that INR values vary linearly between two measurements, although the test results we see from the GP practices may not reflect this. In Chapter 3.4, patients were classified as exposed to anticoagulants based on INR measurements. There is the potential for exposure misclassification given some patients will have been erroneously assumed unexposed. If INR is measured in hospital or at a specialist clinic, data recording relies upon feedback to the GP practice. Missing information on INR measurements is likely to be clustered by practice, where some practices hold anticoagulation clinics in house and others send their patients to be measured elsewhere, such as hospital outpatient departments. If the results are not electronically linked to the GP practice, it may be that only results of interest are fed back, for example those that are outside of therapeutic range. Linked data were not able to add any further information on INR

measurements and therefore it is possible that these studies suffered from some form of exposure misclassification.

Overall, chapter 3 demonstrated the use of data from CPRD to study mortality for a range of conditions. For some conditions, using a primary care database may limit the population by missing those who only have secondary care records. This can be supported by the use of linked secondary care data as shown in chapter 3.3. However, the main limitation with using linked data is the reduction in cohort size since these data are only available for a subset of the CPRD primary care population. This can result in a shorter period of follow-up time, as seen in chapter 3.2, and potentially a loss of precision as seen in chapter 3.4.

Aim 3: To assess the impact on the study population size and follow-up time when using linked data for observational research studies

Studies using observational data from primary care may be missing relevant information from secondary care if this is not relayed to the GP and recorded in the patient record. Many conditions are treated across primary and secondary care; looking at only one data source could leave out considerable information. A previous study examined CPRD primary care data, HES and ONS death registration data to describe the epidemiology of disease across primary and secondary care, however the data sources were not linked together and the results for each data source were presented separately [32]. The availability of individual level data linkage provides the opportunity to validate exposures and outcomes, potentially reducing misclassification in pharmacoepidemiology studies. However, the introduction of linked data to the field of pharmacoepidemiology brings new challenges in terms of the impact of the choice of data source on the generalisability of the findings based on the healthcare setting and the challenges in the definition of the denominator and follow-up time.

The third aim was to assess the impact of using linked data in terms of the decisions made on the definition of the denominator and the calculation of follow-up time. When using linked data, it is important to understand the eligible population and consider this in the definition of the study denominator. Chapter 4 compared the results for the same question when using primary care data only, linked data, and with variations in the calculation of the denominator. The study repeated the analyses from chapter 3.3, examining all-cause mortality following VTE for the period 2005-2009. Four cohorts were identified in order to assess the impact of the choice of data source, and two further cohorts to assess the influence of the denominator (Table 1). The largest cohort was defined using all available data over the widest study period, with the smallest cohort based on patients with a record of VTE from secondary care. Mortality rates varied depending on the choice of data source for the exposure and the outcome, and on the definition of the coverage period and linkage eligibility.

Influence of the healthcare setting

The mortality rate reported in the primary care cohort (108.4/1000 person years) was comparable to a previous published study (97.8/1000 person years) [33], however there are limitations to studying a condition such as VTE solely in primary care. The addition of hospital records identified further VTE cases that had not been reported in primary care. Only 25% of the patients identified in HES had a matching record in primary care on the same date, suggesting gaps in the feedback to practices surrounding events from hospital. This fits with previous studies, where events recorded in hospital were not always reflected in the GP record [34]. Patients identified in secondary care had a higher mortality rate than all other cohorts. This

suggests that those presenting in secondary care may have more severe VTE events (such as pulmonary embolism) or are generally more unwell, at a higher risk of death. The mortality rate for the linked exposure cohort was between the two, reflecting the combined nature of the population with VTE identified from either source. This suggests that for conditions where presentation and management are likely to be shared between primary and secondary care, using both data sources is important to more accurately reflect the patients in actual clinical care.

This study highlighted that consideration must be taken in deciding whether to use linked data in terms of which healthcare setting is most appropriate in order to answer the research question using a representative sample of the target population. CPRD primary care data may not be the best choice to study secondary care outcomes or conditions managed in hospital. However, if the cohort is defined based on secondary care events, this may not provide a true picture of the mortality rates for all patients presenting with this condition. Where additional linked data may support a study, researchers need to consider which data source is best for identifying the population of interest.

Impact on the denominator and follow-up time

Researchers need to consider individual eligibility for linkage and the definition of the start and end of follow-up time when preparing the study denominator. In chapter 4, two cohorts were created solely to assess the impact of making mistakes when preparing the denominator (All Data and Consenting practices cohorts, see Table 1). The CPRD linkage programme is currently restricted to practices in England. Within these practices, individual patients can opt out from their information being used for linkage and some patients have missing information on the NHS number. This means that even within a consenting practice, not all of the registered patients are eligible for each linkage. CPRD have provided a flag to assist researchers in identifying the eligible patients and advise that researchers only include those patients in their linked research studies.

In deciding whether to include linked data, it is important to consider the impact on cohort size. Restricting the population to those from consenting practices and who are eligible for linkage can greatly restrict the cohort size (39% in chapter 3.2; 41% in chapter 3.3; 39% in chapter 3.4) and follow-up time (2.2 versus 1.7 years, chapter 3.2). However, the limitations on cohort size and follow-up time have lessened over calendar time. As of May 2019, the proportion of patients eligible for linkage has increased to 54% [35]. The continuous data collection and recruitment of practices into the linkage programme at CPRD means that the amount of follow-up time available for a research study conducted in primary care data in 2011 (chapter 3.1; patients diagnosed between 1991 and 2007, 1.7 years median follow-up) is comparable to a study using linked data in 2014 (chapter 3.2; patients diagnosed

between 2005 and 2010, 1.7 years median follow-up), despite limiting the cohort to the 39% eligible for linkage in the latter.

In addition to considering individual eligibility, researchers need to consider the definition of follow-up time. Studies based in primary care only can determine the start of follow-up using the point of registration at the GP practice or the date of a diagnosis or prescription, as appropriate. Introducing a second or third data set with varying coverage periods adds complexity to the inclusion of study time. Where the coverage periods of the data sources do not overlap there is the potential to analyse periods of time for which records could not be collated. CPRD primary care data began in 1987, HES data are available from 1997, ONS death registration data from 1998. Previous research using the linked CPRD data may have misunderstood the eligibility criteria or miscalculated the denominator and this may have impacted the results of these studies. Where linked data are available, researchers are likely to use both primary and secondary care data sources to identify the cohort of patients. However, simply pooling data from all sources may introduce errors that could overestimate or underestimate study rates. The all data cohort in chapter 4 had the largest number of patients, however this design would have included follow-up time during which time there was no linked death data (1995-1997). In this case, including all patients, regardless of eligibility or coverage time inflated the denominator, reducing the corresponding mortality rate despite identifying the most deaths (Table 1).

For a study of newly diagnosed multiple sclerosis (MS) patients, the primary care data collection started in 1987 and ended in August 2009 and HES data collection started in April 1997 and ended in March 2008. The authors defined the total study period as 1987 to 2009 and patients were included if they had a diagnosis of MS during the period of GPRD or HES data collection [36]. With hindsight, this would have been done differently (personal communication). For this study, the authors were not able to identify MS diagnoses that were made in hospital between 1987 and 1997, or between 2008 and 2009. Therefore, some of the incident MS patients may not have been newly diagnosed. The denominator will have been inflated to include a wider coverage period, during which time outcomes could not have been ascertained, such as fracture events recorded in hospital in 2009. In chapter 4, a similar cohort was created, mimicking this design. The All Data cohort had the largest study population and most person time of follow-up. However, patients with a VTE recorded in hospital between 1995 and 1997 could not be included and death certificates from before 1998 would not have been picked up. This explains why the resulting mortality rate was the lowest for all cohorts in the study. This design may be comparable with the MS paper, suggesting a correction of the denominator in the study may have resulted in higher fracture rates [36].

For a study of long-acting beta 2-agonists in adult asthma patients, de Vries *et al.* (2010) [37] included questionable follow-up time for analyses of HES data. The study population was based on a cohort identified from primary care prescriptions from January 1993, with outcomes identified in HES. At the time of the study, HES data were available from 2001 to 2007, therefore it would not have been possible to assess hospital-based outcomes between 1993 and 2001. These outcomes were limited to the 200 practices that had consented to linkage, a small proportion of the original cohort. It is not clear from the publication how the denominators for each outcome were defined, whether the coverage period was limited to the overlapping time or if the authors assessed individual eligibility for linkage. This is likely to have had an impact on the denominator and in turn the reported incidence rates of hospitalisation. In chapter 4, the consenting practices cohort used a similar design. Patients from practices that had not signed up to the linkage programme were excluded, however individual patients were included regardless of their eligibility for linkage. This would have resulted in the inclusion of VTE patients from primary care who could not have had a death recorded in the ONS data; the mortality rate was therefore lower than we might have otherwise expected. In chapters 3.2, 3.3 and 3.4, additional consideration was required in the definition of the study window and coverage periods to avoid the issues described. Time was spent reviewing which practices had consented to take part in the linkage programme, which patients registered at those practices were eligible based on their identifiers, and the time periods of availability of the linked data. Once the denominator and inclusion and exclusion criteria had been applied, discordance between the data sources also had to be considered. This was particularly important where an alternative death date affected the censoring point for analyses, where some patients were then excluded from the study. In the study population analysed in chapter 3.3 there were 12 patients who died on the index date and 24 who died before the start date for the study, all of these were excluded. In addition, there were 2,161 patients with a death date recorded in the ONS data that was after the censoring date from the practice information. This information was therefore not used in the analyses. In the linked exposure and outcome cohort in chapter 4, for patients with a death record in both data sources, there were 862 patients (18.3%) where the date in the death certificate was earlier than recorded in primary care. This equated to 50 person years of follow-up time which would be included for studies in the primary care setting for patients who, based on the ONS data, had already died.

Since there was some discordance between the death date identified in the primary care records, using the CPRD algorithm and the date on the death certificate, further analyses were performed using the linked exposure and linked outcome cohort from chapter 4 to assess sensitivity and specificity with regard to the CPRD algorithm (see Table 2, data not shown in chapter 4). For the population eligible for linkage with both HES and ONS data, with VTE recorded in either primary or secondary care (Linked

exposure and outcome cohort, N=22,639), 5,539 patients had a death date record in CPRD within the two years follow-up time after diagnosis. Of these, 4,719 also had a death certificate, presenting a positive predictive value (PPV) of 85.2%. There were just 30 patients that were assumed alive in the primary care record but that had a death certificate, presenting a sensitivity of 99.4%. The high sensitivity, coupled with a relatively low PPV, suggests that the algorithm used by CPRD to identify deaths is more likely to erroneously report a patient death than to miss patients who have truly died. The impact would be different depending on the nature of the study; where death events are used as a censoring point, these patients would be censored earlier than they should be, however, they may be misclassified as dead in a study of all-cause mortality.

Chapter 4 looked at how using linked data can increase misclassification depending on how the denominator is applied. The findings show that it may be easy to make a mistake with censoring, however there may be an impact on the results of a study. In conclusion, it is important to get the denominator right to avoid misclassification and inflation of person time.

Table 1: Cohort descriptions from chapter 4

	Linkage eligibility	Coverage period	Exposure (VTE)	Outcome (death)	Number of patients	Number of deaths	Mortality rate per 1000 person years
Primary care cohort	N/A	1995-2009	Primary care	Primary care	36,216	6,112	108.4
Linked exposure cohort	Eligible for linkage with hospital records	1997-2009	Primary care or hospital records	Primary care	23,720	5,771	169.9
Linked exposure and outcome cohort	Eligible for linkage with hospital records or death certificates	1998-2009	Primary care or hospital records	Primary care or death certificates	22,639	5,569	173.2
				Primary care	22,639	5,539	171.9
				Death certificates	22,639	4,749	147.3
Secondary care cohort	Eligible for linkage with hospital records	1997-2009	Hospital records	Primary care	13,404	4,186	237.2
All data cohort	Any	1995-2009	Primary care or hospital records	Primary care or death certificates	46,296	9,701	140.6
Consenting practices cohort	Patients registered at consenting practices	1995-2009	Primary care or hospital records	Primary care or death certificates	26,320	6,264	164.3

VTE: Venous Thromboembolism

Table 2: Death records for patients in the linked exposure and linked outcome cohort from chapter 4 (N=22,639)

		Death recorded in the ONS death registration data			
		Yes	No		
Death recorded in the CPRD	Yes	4,719	820	5,539	*Positive predictive value* = 85.2%
	No	30	17,070	17,100	*Negative predictive value* = 99.8%
		4,749	17,890		
		Sensitivity = 99.4%	*Specificity* = 95.4%		

CPRD: Clinical Practice Research Datalink; ONS: Office for National Statistics

Aim 4: To evaluate the agreement between English primary care data and national death registration data in capturing vital status and the date of death

In chapter 5 the aim was to compare the data in CPRD to national statistics. The ONS death registration data have been assumed to be the gold-standard for records on the vital status, date and cause(s) of death. However, since linked death data has been available for research, not all studies have chosen to use it. Some CPRD studies with a mortality outcome were designed purely in primary care, without using any linked data [38-41]. Other studies have used linked HES data in the definition of their exposure, covariates or outcomes, thereby restricting the cohort to linkage-eligible patients and then assessed all-cause mortality without using the available linked ONS data [42-44]. This suggests either confidence in the primary care record/CPRD algorithm in recording the fact and date of death, despite a lack of published information on the accuracy of the death information in CPRD, as reported in one study "Mortality was comprehensively ascertained from dates of death recorded into primary care records" [39] or uncertainties in the data from the death certificates. The ONS death registration data are considered the gold standard for mortality statistics, however it is important to understand how and why these data are collected and recorded in order to make judgements on the likely accuracy.

The accuracy of the national death registration data

Databases of death certificates have been considered the gold-standard for information on the date and cause of death. Although not originally intended for research, death certificates provide useful information for epidemiological studies. Death certificates provide more than one function; they have a legal and administrative function as well as providing vital statistics for epidemiology and health policy/planning purposes. If researchers are to consider using this data source for research studies, they first need to consider the accuracy of the records and any caveats in the data recording. Since 1837, it has been a legal requirement that all deaths in England and Wales are registered with the Local Registration Service, in partnership with the General Register Office (ONS, 2018). Death registration details, including the date and cause of death, are recorded on the Registration Online system by registrars, based on information supplied by the informant (usually a close relative of the deceased). The cause of death is usually obtained from the Medical Certificate of Cause of Death (MCCD), which will have been completed by a medical practitioner when the death is certified [45].

Despite the importance of accurate death certification, errors are common, particularly in the recording of the cause of death, where the concept of 'underlying cause of death' is often a source of confusion for the certifier. Although most clinicians will be faced with the task of completing a death certificate, many will not receive adequate training and there is some uncertainty about the proper process for

completion [46, 47]. Guidelines have been updated over the years but mainly pertain to the recording of the cause of death [48-50]. There is a lack of importance placed on teaching death certification, with calls to modernise this during medical training [51]. Introducing a tutorial to teach medical students how to correctly complete a death certificate at one medical center in the US was found to be effective [52] and an educational intervention in Ontario reduced the major errors in the cause of death statement from 32.9% to 15.7% [53]. However, in Scotland it was noted that the six medical reviewers would not be able to deliver an education programme to around 12,000 doctors and 1,000 new graduates every year [54].

The system for death certification in England has remained largely unchanged for over 50 years. Death certification reform was initiated following the conviction in 2000 of Harold Shipman, an English family doctor who murdered 15 of his patients. An inquiry found that he may have murdered many more patients, certifying the deaths as natural. Through a phased roll-out from April 2019, medical examiners will be introduced to the death certification process to confirm the cause of all deaths that do not need to be investigated by a coroner. Once fully rolled out the system will prevent burials being authorised on the unchecked certificate of the treating doctor [55].

Concerns about inaccuracies in death certificates are long-standing and are not confined to the United Kingdom. A study in Iceland found significant discrepancies for 50% of patients [56] and 96% contained some degree of error in a study in Vermont [57]. In advance of establishing a new death certification review service, a pilot study in Dundee found 1 in 16 death certificates contained errors. Mistakes included inaccurate causes of death, errors in the sequences of events leading up to a death, the wrong death date or even wrong name. A previous study investigated the accuracy of data in death certificates in England and Wales for a population of patients implanted with Bjork-Shiley Convexo-Concave (BSCC) heart valves [58]. The authors assessed whether the death certificate included information on the prosthesis, the place of death, the occurrence of a post-mortem and whether the mode of death rather than the cause was recorded. They concluded that research studies based solely on the review of death certificates, without clinical case note review, may lead to presentation of inaccurate results and possible underreporting of significant events.

In England and Wales, deaths should be registered within five days of the date of death. This fast turnaround time is supportive of research however the death can only be registered once there is a certified cause of death. Sudden infant deaths, drug related deaths, suicides, deaths that occur in prison or police custody and unexplained deaths will be referred to a coroner who may order a post-mortem or carry out a full inquest to establish the cause; this can take between 6 months and two and a half years, during which time the death is not recorded in the national

statistics for England and Wales. Around 10,000 deaths a year in England and Wales are not registered for at least 6 months [59]. For the causes most affected by late registrations (suicides and drug-related deaths), around half occur outside their registration year [60]. This has consequences for research. A study of treated opiate users in England was based on data from drug treatment and criminal justice sources for the period 2005–2009. In order to be confident in knowing about all deaths occurring before April 1st, 2009, Pierce *et al*. (2013) [61] had to delay their check against the deaths register by two and a half years, until October 2011 to allowed for late registration of deaths referred for investigation by a coroner.

The ONS death registration data are based on information recorded when deaths are certified and registered. They include only deaths that occur in England and Wales and not those that occur in other parts of the United Kingdom or abroad. The information normally comes from 1 of 3 sources, for which completeness may vary [62]:

1. The doctor who certified death
   for example: cause of death, whether a post-mortem was carried out
2. The informant to the registrar
   for example: sex, usual address, date and place of birth, date of death, place of death
3. The coroner to the registrar following investigation
   for example: cause of death (following post-mortem), place of accident (following inquest)

With all this in mind the findings of studies based on death certificates need to be interpreted with caution, particularly for certain causes of death and for time-sensitive studies where waiting for up to two and half years to ensure the records are complete is not feasible.

Despite the known shortcomings, information from death certificates continue to be used for national statistics and for research projects evaluating public health interventions. Death certification reform in England is likely to improve the accuracy of the records, however they are still considered the gold-standard for vital status, cause of death information and accurate recording of the date of death.

The agreement between English primary care data and national death registration data

Chapter 5 focused on the agreement in the recording of deaths between data sources. There was substantial agreement between the national records from the ONS and the data from CPRD and almost half of all deaths were recorded in the HES. The comparison was based on individually linked data and it is the discordance that is of most interest. Some deaths were recorded as part of the hospital audit record or in the primary care record without being reflected in the national statistics. This may reflect the delays in registration as previously discussed.

The mortality rate based on HES data was around half of that in primary care or the ONS, which fits with published figures that as many as half of deaths occur in hospital. The paper did not recommend using the HES as the sole source of data for identifying death as an outcome, however for a study of hospital deaths, the HES data may be a good source for identifying information on vital status.

This study highlighted that some deaths are not recorded at the GP practice. There is no link between the GP practice and the official certification of death; accurate recording relies upon information being passed to the practice and it being entered into the patient record. A large proportion of deaths occur in hospitals, hospices, nursing and care homes, and it is also no longer a requirement for the local GP to certify a death in person, even when this occurs in the community. Therefore for most deaths, the GP practice will be notified from an external source. In a study over 15 years in a single UK practice, 86.2% of deaths were formally notified to the practice, mostly from hospitals, coroners and hospices. Informal notification of death (from a relative, neighbour or newspaper report) was the only information received for 13.8% of deaths. For 67.8%, the notification was within 7 days of death, however, for 15% the practice was never notified [63].

Missing information for some patients on the fact of death and/or the date of death would impact on the inclusion of patients in a study, the denominator definition and on the rates of all-cause mortality. For a case control study where the population is identified based on a death outcome, missing information on vital status in the primary care record would mean that some individuals are excluded from the study. Conversely, patients may be included in the study with a death record in primary care that has not been confirmed by the ONS. This study design has previously been used to examine individuals aged 40 or more, who were newly prescribed antidiabetic drugs, where cases were all those who had a primary care death record [64]. The high sensitivity of the CPRD algorithm found in the analyses of chapter 4 suggest most deaths are reported to the practice, however the relatively low PPV suggests we cannot be sure of the true vital status of all those with a death record in primary care. For cohort studies, missing information on the date of death could

mean that some patients contribute follow-up time after their death. This additional follow-up time would inflate the denominator and the calculation of all-cause mortality could be underestimated due to missing information in the numerator. Both case control and cohort study designs are common in pharmacoepidemiology.

The expectation is that once a practice is informed about a death, the patient's record is updated, with a backdated reporting of death to match the death certificate. However according to the British Medical Association, GPs only have to report cause and not the fact of death [65]. The date recorded may reflect the date the healthcare professional heard about the death, closed the patient's record, or ended an episode of care rather than the actual date of death. It is important to note that GPs are under time pressure; the burden on the primary care physician has increased [66] and it may not be so important for the GP to record death information well given further clinical care is not required. Despite this, analyses in chapter 5 found that death records at the GP practice were predominantly recorded with the same date as on the official death certificate (68.4%). However, for 27.2% of deaths, the date of death was noted as later than recorded on the death certificate. The delay was usually less than 30 days and is likely to reflect the current date at the point of notification. For some deaths the difference between the dates was exactly a year, suggesting a typo in the recording of the year, such misdating may be more common in the early months of a calendar year.

Harshfield *et al.* (2018) [67] conducted similar analyses to those presented in Chapter 5, comparing death records in CPRD to the ONS death registration data. This study was based on the population in CPRD who were aged over 18 with a death record in the primary record between 2010 and 2015. The authors' conclusions were less positive about the agreement between the data sources, despite a higher proportion of exact agreement in their study population (76.8%). Additional comparisons by cause of death found the discrepancies to differ by cause, particularly for those designated 'external' (which includes accidents, assaults and intentional self-harm). This fits with the expectation that these deaths may be affected by late registration and subsequent delayed feedback to the GP practice. Since there is likely to be late registration for some causes of death, some consideration should be made when deciding on the end of the study period for analyses using linked data from the ONS. For pharmacoepidemiology, the adverse event of sudden death is likely to be better captured in the ONS data in comparison to primary care, however a time lag is necessary to ensure that all deaths are registered, which could be as long as 2.5 years. This means these data sources are not ideal for assessing short-term mortality, particularly with new drugs where the population of users may be small.

Chapter 5 compared the data in primary care to that from the national death certificates. The overall level of concordance was good, however given that there may be delays in a practice being informed about deaths and limited incentive to record this accurately, some consideration should be made when using primary care data alone to assess deaths.

Future direction

The analyses in chapter 5 examined the difference between the primary care record and the date of death from the death certificate but not the date of registration. Comparing the date of death from the algorithm with the date of registration of the death would be interesting in order to assess whether the recording at the GP practice is more or less accurate when the death was registered quickly.

Further analyses on the delays in death registrations by cause of death, to complement the study by Harshfield *et al.* (2018) [67], could support studies of cause-specific mortality, where researchers may need to choose the end of their study period based on the likelihood of late registration. In addition, further work could assess the delays in death registrations by place of death. This field has recently been included in the ONS death registration data linked to CPRD. It is possible that there are differences in the recording of deaths depending on whether an individual dies at home, compared to in hospital or in a nursing or care home.

The CPRD algorithm for estimating the data of death was not examined in detail in this thesis. The algorithm is based on four elements, each of which may contribute to a greater or lesser extent. Researchers could benefit from understanding how well the date associated with the 'Transferred out' reason of death matches the ONS records and whether this alone could be a suitable marker for deaths in primary care. In addition, the algorithm currently includes only events occurring within 95 days of the date of transfer out. Further analysis of all patients who have a death certificate could help to ascertain whether this time window is appropriate. Changes in recording behaviour over time may affect the accuracy of the current algorithm and CPRD may want to consider reviewing this using more recent data, particularly as an algorithm is yet to be developed for the new CPRD Aurum database.

The analyses presented were all based on the CPRD GOLD database, which is based on practices using Vision software. CPRD have since released the Aurum database which include longitudinal data from practices using the EMIS software [68]. Although the practice is likely to be recording the same information for the purposes of clinical care, the software systems are different and this may have an impact on the recording of death information. Further work in this field could compare the death records in CPRD Aurum to the ONS death registration data in a similar way to this work for Vision using practices.

Recommendations

There may be two separate reasons why the date of death record at the GP practice is different from the death certificates from the ONS. Firstly, there are likely to be delays in the GP being informed about a death of one of their patients. Secondly, there are increasing pressures on GPs in the UK and there may be limited incentive to accurately record the date of death. Currently, death certificates do not include details of the GP with whom the patient was registered at death. This is likely to impact how soon a GP is informed. Including the GP details as part of upcoming death certification reform may improve the delays in information being passed to the registered GP. There is currently no link between primary and secondary care or with the death registry, at the healthcare professional level and limited training on the completion of death certificates. An electronic system for death certification would have obvious advantages. Enabling GPs to issue death certificates electronically would save GP time as well as inconvenience for bereaved relatives. The system could include guidance for completion of death certificates, and the need for repeated data entry would be reduced. There would also be benefits from having a readily accessible audit trail and less likelihood of typographical errors from retrospective recording.

The ONS may not have complete data on the fact of death until there is a cause established. This delay is more likely to affect sudden or unexpected deaths and any that are reviewed by a coroner. In Scotland the registration of the fact of death has been uncoupled from registration of the cause of death. Without exception, all deaths in Scotland must be registered within 8 days of death having been ascertained, therefore delays are much less marked. Data for England and Wales may benefit from a future uncoupling however, this would require a change in legislation which others have campaigned for [69].

In Scotland, errors have reduced by more than a third since its Death Certification Review Service was established in 2015 [70]. The review service checks on the accuracy of a sample of death certificates, designed to improve accuracy, provide better quality information about causes of death so that health services can be better prepared for the future, and ensure that the processes around death certification are robust and have appropriate safeguards in place. A similar system in England could similarly reduce the errors in death certification.

Conclusion

In conclusion, the information on vital status and the date of death may not be accurately recorded in primary care records. Although concordance with the national death certificate data for England is good, the primary care data will often have a date of death recorded that is later than the date on the death certificate and this will have a larger impact on studies of person time. The hypothesis was that there are sufficient data recording at the GP practice to enable research with this data source and for certain study designs and exposure/outcome pairs, this was adequately met.

Studies based within primary care records are limited in their ability to assess the cause of death and may not be representative of all patients with the condition/exposure of interest. These issues can be addressed by using individually linked data, such as offered by the CPRD in the supply of primary care data linked to secondary care records and death registration data. In using linked data, there are a number of limitations which need to be balanced against the additional benefits of added information available through linkage. The main limitation is the restriction on cohort size, which can be significant and may impact on the available sample size or power of a study. Further limitations include the complexity of working with linked data, particularly in defining the eligible study population and appropriate follow-up time for analyses. Despite the limitations, it is important to remember that these data sources are based on documented, rather than recalled information and the availability of data linked at the individual level represents a significant revolution in the field of pharmacoepidemiology.

Longitudinal data from primary care offer a good source of information for studies evaluating rare outcomes and to assess long latency periods. For adverse events that appear within hours of treatment, deaths are likely to be reviewed by a coroner, in these cases there is likely to be a delay in the registration of the death, however where it is important to ascertain the cause, it would be wise to wait for the available ONS data. Conversely, there may be some situations where the GP record includes information on vital status and the date of death before the release of ONS death registration data, examples include situations where the death can't be registered until the inquest has been completed but the GP has been informed; for studies of all-cause mortality the primary care data may therefore be most useful.

Finally, this thesis shows that although primary care databases will always be based on data recorded for clinical care purposes, there is good recording of vital status and date of death. Linked data has great potential to fill gaps in recording however there are challenges with cohort size, denominators and censoring.

References

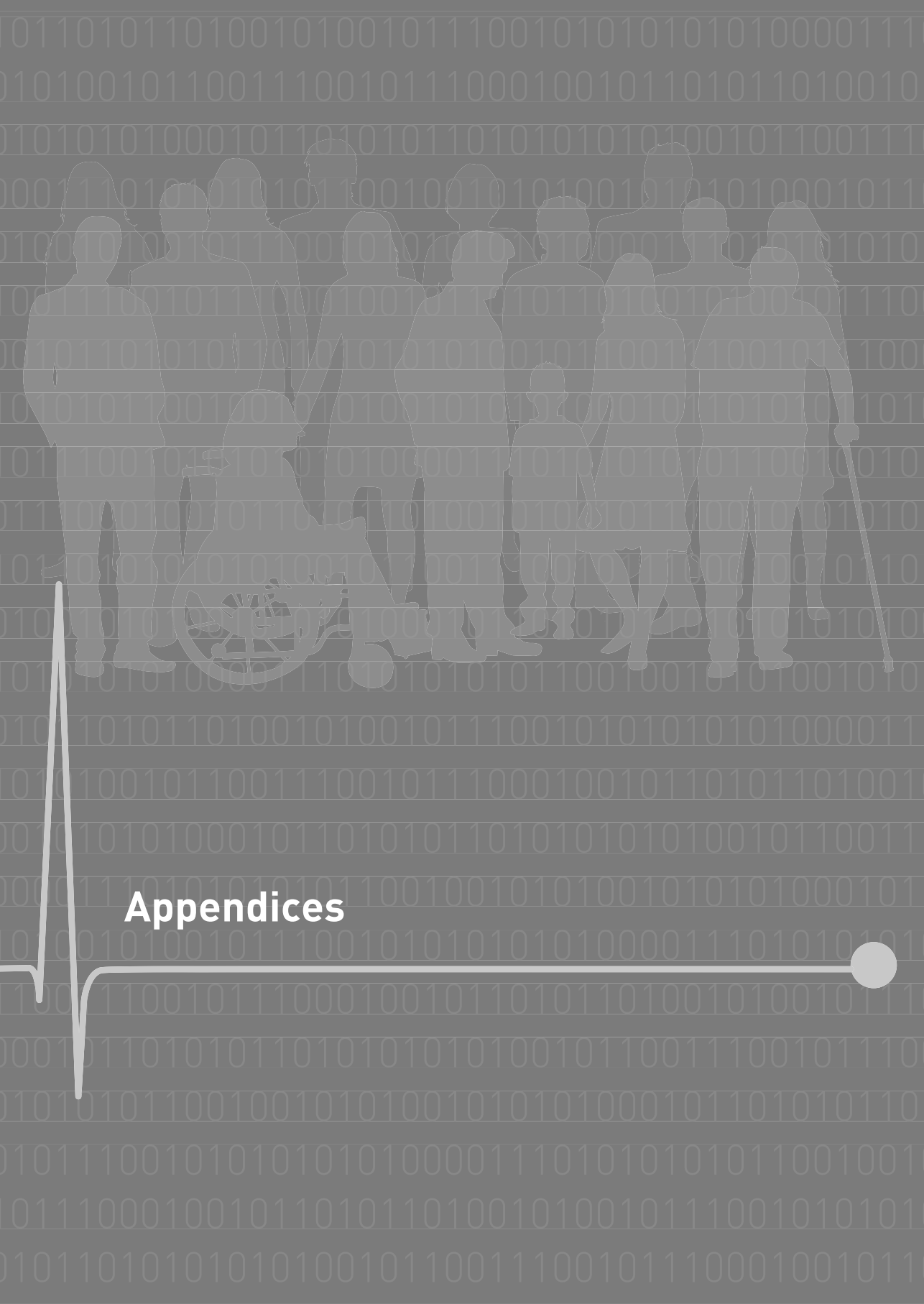
1. Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK primary care data resource. Ther Adv Drug Saf. 2012;3:89-99.
2. Dimensions. Available at: https://app.dimensions.ai/details/publication/pub.1014928761?and_facet_researcher=ur.01124521745.61 [Accessed 29 April 2019]
3. Campbell J, Dedman DJ, Eaton SC, Gallagher AM, Williams TJ. Is the CPRD GOLD Population Comparable to the UK Population? Pharmacoepidemiol Drug Saf. 2013;22:S280-281.
4. Setakis E, Puri S, Williams TJ, VanStaa TP. Representiveness of Subset of the General Practice Research Database (GPRD) Linked to Other Data Sources. Pharmacoepidemiol Drug Saf. 2010;19:S311.
5. Gallagher AM, Puri S, van Staa TP. Linkage of the General Practice Research Database (GPRD) with Other Data Sources. Pharmacoepidemiol Drug Saf. 2011;20:S230-231.
6. Office for National Statistics. Deaths registered by area of usual residence, UK. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsregisteredbyareaofusualresidenceenglandandwales. [Accessed 29 April 2019]
7. Kontopantelis E, Mamas MA, van Marwijk H, Ryan AM, Buchan IE, Ashcroft DM, Doran T. Geographical epidemiology of health and overall deprivation in England, its changes and persistence from 2004 to 2015: a longitudinal spatial population study J Epidemiol Community Health. 2018;72:140-147.
8. Office for National Statistics. Deaths registered in England and Wales: 2017. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/2017 [Accessed 29 April 2019]
9. Marshall T, Simpson S, Stevens A. Use of health services by prison inmates: comparisons with the community. J Epidemiol Community Health. 2001;55:364-5
10. Feron JM, Paulus D, Tonglet R, Lorant V, Pestiaux D. Substantial use of primary health care by prisoners: epidemiological description and possible explanations. J Epidemiol Community Health. 2005;59:651-5
11. Sturge G. UK Prison Population Statistics. Briefing Paper Number CBP-04334, 23 July 2018. House of Commons Library.
12. GOV.UK. Official Statistics: Prison population figures: 2019. Available at: https://www.gov.uk/government/statistics/prison-population-figures-2019. [Accessed 7 May 2019]
13. NHS. Armed forces healthcare: how it works (2018). Available at: https://www.nhs.uk/using-the-nhs/military-healthcare/armed-forces-healthcare-how-it-works/. [Accessed 3 May 2019]
14. Ministry of Defence. Deaths in the UK regular armed forces: 2018. Available at: https://www.gov.uk/government/statistics/uk-armed-forces-deaths-in-service-2018. [Accessed 3 May 2019]
15. GOV.UK. National Statistics: Quarterly service personnel statistics: 2018. Available at: https://www.gov.uk/government/statistics/quarterly-service-personnel-statistics-2018. [Accessed 7 May 2019]

16. Martinez-Gomez D, Guallar-Castillon P, Higueras-Fresnillo S, Banegas JR, Sadarangani KP, Rodriguez-Artalejo F. A healthy lifestyle attenuates the effect of polypharmacy on total and cardiovascular mortality: a national prospective cohort study. Sci Rep. 2018;8:12615
17. Dwyer LL, Han B, Woodwell DA, Rechtsteiner EA. Polypharmacy in nursing home residents in the United States: Results of the 2004 National Nursing Home Survey. Am J Geriatr Pharmacother. 2010;8:63-72
18. Marengoni A, Onder G. Guidelines, polypharmacy, and drug-drug interactions in patients with multimorbidity. BMJ. 2015;350:h1059.
19. National End of Life Care Intelligence Network. Deaths in older adults in England. Available at: http://www.endoflifecare-intelligence.org.uk/resources/publications/deaths_in_older_adults. [Accessed 3 May 2019]
20. van Dam RM, Li T, Spiegelman D, Franco OH, Hu FB. Combined impact of lifestyle factors on mortality: prospective cohort study in US women. BMJ. 2008;337:a1440.
21. Khaw KT, Wareham N, Bingham S, Welch A, Luben R, Day N. Combined impact of health behaviours and mortality in men and women: the EPIC-Norfolk prospective population study. PLoS Med. 2008;5:e12.
22. Petersen I, Welch CA, Nazareth I, Walters K, Marston L, Morris RW, Carpenter JR, Morris TP, Pham TM. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. Clin Epidemiol. 2019;11:157-167.
23. Crane J, Pearce N, Burgess C, Woodman K, Robson B, Beasley R. Markers of risk of asthma death or readmission in the 12 months following a hospital admission for asthma. Int J Epidemiol. 1992;21:737-44.
24. Jones M, McEwan P, Morgan CL, Peters JR, Goodfellow J, Currie CJ. Evaluation of the pattern of treatment, level of anticoagulation control, and outcome of treatment with warfarin in patients with non-valvar atrial fibrillation: a record linkage study in a large British population. Heart. 2005;91:472-7.
25. Currie CJ, Jones M, Goodfellow J, McEwan P, Morgan CL, Emmas C, Peters JR. Evaluation of survival and ischaemic and thromboembolic event rates in patients with non-valvar atrial fibrillation in the general population when treated and untreated with warfarin. Heart. 2006;92:196-200.
26. Morgan CL, McEwan P, Tukiendorf A, Robinson PA, Clemens A, Plumb JM. Warfarin treatment in patients with atrial fibrillation: Observing outcomes associated with varying levels of INR control. Thromb Res. 2009;124:37-41.
27. Johansson C, Dahlqvist E, Andersson J, Jansson JH, Johansson L. Incidence, type of atrial fibrillation and risk factors for stroke: a population-based cohort study. Clin Epidemiol. 2017;9:53–62.
28. Ruigómez A, Johansson S, Wallander MA, Rodríguez LA. Incidence of chronic atrial fibrillation in general practice and its treatment pattern. J Clin Epidemiol. 2002;55:358-63.
29. Lawrenson R, Todd JC, Leydon GM, Williams TJ, Farmer RD. Validation of the diagnosis of venous thromboembolism in general practice database studies. Br J Clin Pharmacol. 2000;49:591-6.
30. Rosendaal FR, Cannegieter SC, Van der Meer FJM, Briët E. A method to determine the optimal intensity of oral anticoagulant therapy. Thromb Haemost 1993;69:236-9.
31. Hutten BA, Prins MH, Redekop WK, Tijssen JG, Heisterkamp SH, Büller HR. Comparison of three methods to assess therapeutic quality control of treatment with vitamin K antagonists. Thromb Haemost. 1999;82:1260-3.

32. Gupta D, Hansell A, Nichols T, Duong T, Ayres JG, Strachan D. Epidemiology of pneumothorax in England. Thorax. 2000;55:666-71.
33. Huerta C, Johansson S, Wallander MA, Rodríguez LA. Risk of myocardial infarction and overall mortality in survivors of venous thromboembolism. Thromb J. 2008;6:10.
34. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, Timmis A, Hemingway H. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013;346:f2350.
35. CPRD. CPRD linked data. Available at: https://cprd.com/linked-data [Accessed 22 June 2019]
36. Bazelier MT, van Staa T, Uitdehaag BM, Cooper C, Leufkens HG, Vestergaard P, Bentzen J, de Vries F. The risk of fracture in patients with multiple sclerosis: The UK general practice research database. Journal of Bone and Mineral Research. 2011;26:2271–9.
37. de Vries F, Setakis E, Zhang B, van Staa TP. Long-acting {beta}2-agonists in adult asthma and the pattern of risk of death and severe asthma outcomes: a study using the GPRD. Eur Respir J. 2010;36:494-502.
38. Alatorre CI, Hoogwerf BJ, Deeg MA, Nelson DR, Hunter TM, Ng WT, Rekhter MD. Factors associated with stroke, myocardial infarction, ischemic heart disease, unstable angina, or mortality in patients from real world clinical practice with newly-diagnosed type 2 diabetes and early glycemic control. Curr Med Res Opin. 2018;34:337-343.
39. Charlton J, Ravindrarajah R, Hamada S, Jackson SH, Gulliford MC. Trajectory of Total Cholesterol in the Last Years of Life Over Age 80 Years: Cohort Study of 99,758 Participants. J Gerontol A Biol Sci Med Sci. 2018;73:1083-9.
40. Springate DA, Parisi R, Kontopantelis E, Reeves D, Griffiths CE, Ashcroft DM. Incidence, prevalence and mortality of patients with psoriasis: a U.K. population-based cohort study. Br J Dermatol. 2017;176:650-8.
41. Zghebi SS, Steinke DT, Carr MJ, Rutter MK, Emsley RA, Ashcroft DM. Examining trends in type 2 diabetes incidence, prevalence and mortality in the UK between 2004 and 2014. Diabetes Obes Metab. 2017;19:1537-45.
42. Browne J, Edwards DA, Rhodes KM, Brimicombe DJ, Payne RA. Association of comorbidity and health service usage among patients with dementia in the UK: a population-based study. BMJ Open. 2017;7:e012546.
43. Holden SE, Jenkins-Jones S, Morgan CL, Peters JR, Schernthaner G, Currie CJ. Prevalence, glucose control and relative survival of people with Type 2 diabetes in the UK from 1991 to 2013. Diabet Med. 2017;34:770-80.
44. Wilson JC, Sarsour K, Collinson N, Tuckwell K, Musselman D, Klearman M, Napalkov P, Jick SS, Stone JH, Meier CR. Serious adverse effects associated with glucocorticoid therapy in patients with giant cell arteritis (GCA): A nested case-control analysis. Semin Arthritis Rheum. 2017;46:819-27.
45. ONS 2018. Mortality statistics in England and Wales QMI. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/mortalitystatisticsinenglandandwalesqmi [Accessed 6 June 2019]
46. Maudsley G, Williams EM. "Inaccuracy' in death certification--where are we now? J Public Health Med. 1996;18:59-66.
47. O'Donovan BG, Armstrong P, Byrne MC, Murphy AW; Donegal Specialist Training Programme in General Practice. A mixed-methods prospective study of death certification in general practice. Fam Pract. 2010;27:351-5.

48. Logan WPD. Refresher course for General Practitioners: Death Certification. BMJ. 1953;1272-4.
49. Klatt EC, Noguchi TT. Death certification. Purposes, procedures, and pitfalls. West. J. Med. 1989;151:345-7.
50. Crowcroft N, Majeed A. Improving the certification of death and the usefulness of routine mortality statistics. Clin Med (Lond). 2001;1:122-5.
51. Khan AA, Ah-Kee EY. Death Certification: 800 years of practice; time to modernise teaching? Scott Med J. 2016;61:32-3.
52. Degani AT, Patel RM, Smith BE, Grimsley E. The effect of student training on accuracy of completion of death certificates. Med Educ Online. 2009;14:17.
53. Myers KA, Farquhar DR. Improving the accuracy of death certification. CMAJ. 1998;158:1317-23.
54. The Scottish Parliament. Certification of Death (Scotland) Bill 2010. Available at: http://www.parliament.scot/S3_Bills/Certification%20of%20Death%20(Scotland)%20Bill/BBV164_-_Final.pdf [Accessed 11 May 2019]
55. Luce T, Smith J. Death certification reform in England. BMJ. 2018;361:k2668.
56. Nielsen GP, Björnsson J, Jonasson JG. The accuracy of death certificates. Implications for health statistics. Virchows Arch A Pathol Anat Histopathol. 1991;419:143-6.
57. Pritt BS, Hardin NJ, Richmond JA, Shapiro SL. Death certification errors at an academic institution. Arch. Pathol. Lab. Med. 2005;129:1476-9
58. Morton L, Omar R, Carroll S, Beirne M, Halliday D, Taylor KM. Incomplete and inaccurate death certification--the impact on research. J Public Health Med. 2000;22:133-7.
59. Parliamentary questions by Patrick Mercer OBE. Available at: https://publications.parliament.uk/pa/cm201212/cmhansrd/cm120313/text/120313w0002.htm#12031386001816. [Accessed 7 May 2019]
60. Office for National Statistics. Deaths related to drug poisoning in England and Wales: 2017 registrations. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsrelatedtodrugpoisoninginenglandandwales/2017registrations. [Accessed 7 May 2019]
61. Pierce M, Bird SM, Hickman M, Millar T. National record linkage study of mortality for a large cohort of opioid users ascertained by drug treatment or criminal justice sources in England, 2005-2009. Drug Alcohol Depend. 2015;146:17-23.
62. Office for National Statistics. User guide to mortality statistics, July 2017. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017 [Accessed 15 February 2018]
63. Beaumont B, Hurwitz B. Is it possible and worth keeping track of deaths within general practice? Results of a 15 year observational study. Qual Saf Health Care. 2003;12:337-42.
64. Azoulay L, Schneider-Lindner V, Dell'aniello S, Schiffrin A, Suissa S. Combination therapy with sulfonylureas and metformin and the prevention of death in type 2 diabetes: a nested case-control study. Pharmacoepidemiol Drug Saf. 2010;19:335-42.
65. BMA. Confirmation and certification of death. Available at: https://www.bma.org.uk/advice/employment/gp-practices/service-provision/confirmation-and-certification-of-death [Accessed 9 March 2019]
66. O'Malley PG, Greenland P. The annual physical: are physicians and patients telling us something? Arch Intern Med. 2005;165:1333-4.
67. Harshfield A, Abel GA, Barclay S, Payne RA. Do GPs accurately record date of death? A UK observational analysis. BMJ Support Palliat Care. 2018;0:1-8.

68. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, Myles P. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol. 2019;pii: dyz034.
69. Bird SM. Counting the dead: properly and promptly. Lancet. 2013;381:1719.
70. Healthcare Improvement Scotland. 2018. Healthcare Improvement Scotland reports a significant improvement in how doctors are recording information on deaths. Available at: http://www.healthcareimprovementscotland.org/news_and_events/news/news_dcrs_ar_2017-18.aspx?theme=mobile [Accessed 11 May 2019]



Appendices

English Summary

One of the most relevant analysed end points in pharmacoepidemiology is death. For public health studies this is often a definitive outcome but death also an important censoring end point in studies with any outcome. For observational research studies, the ability to accurately determine follow-up time is pivotal in epidemiological evaluations and therefore records on vital status (dead versus alive) and the associated date are crucial.

There has been a surge in the use of electronic healthcare records (EHR) for research over the past 30 years. A commonly used EHR database is the Clinical Practice Research Datalink (CPRD) GOLD, based on primary care records from general practices in the UK. This data source is commonly used for pharmacoepidemiology, however there are limitations to the data as the records are primarily collected for the purposes of clinical care or audit. The recording of deaths in UK primary care records is subject to question since there is no direct link between the general practice and the national registry of deaths, and there may be limited incentive for correct recording of vital status and the date of death in the record.

Chapter 1 outlines the available data sources for observational studies of pharmacoepidemiology in Europe, with a focus on the UK primary care records in the CPRD GOLD database. This is an internationally recognised data source used for research studies of public health benefit. The database of electronic health records is based on the UK population and relies upon the free at the point of care access provided by the National Health Service (NHS). The CPRD GOLD database is commonly used for pharmacoepidemiology studies however, at the time of starting this thesis there had not been any studies that assessed the accuracy of the status of death and the recorded date in this data source.

It is not clear whether deaths are well recorded in the CPRD GOLD as General Practitioners (GPs) may not routinely receive information about their deceased patients. In the UK, there is no direct linkage between the GP practice and the national registry of deaths, and since primary care electronic healthcare records are based on information originally recorded for the purposes of clinical care, there may be limited incentive for accurate recording of vital status and the date of death. In addition, changes in the software or in recording practices may have an impact on death recording over time.

There are multiple ways of recording a death in the Vision software system used by GPs contributing data to CPRD GOLD. An algorithm was developed to combine recorded information and identify the best estimate for the date of death. This produces a single derived date of death which is added to the database. However, the methodology for the algorithm has not been published and the derived dates have not been validated.

Missing information on vital status is likely to have an impact on the robustness of a study as this could result in the erroneous inclusion or exclusion of individuals in the study population and/or contribute to a miscalculation of person time. The inclusion of follow-up time after an individual has died inflates the denominator for a study and this type of information bias cannot be corrected for with analytic techniques.

Individual level linkage is now available between CPRD GOLD and external data sources such as the Office for National Statistics (ONS) death registration data and the Hospital Episode Statistics (HES) admitted patient care data. These data sources offer the opportunity to add supplementary information on diagnoses, identify additional patients for a study and validate primary care death records. However, linked data are only available for a subset of the CPRD population, based on consenting GP practices in England. This restriction reduces the potential cohort size for a study and adds complexity to the generation of the appropriate denominator.

This thesis aimed to evaluate the utility of the CPRD GOLD database for studies with death as an outcome. The overall thesis objective was to evaluate the use of CPRD data for analysing death information in pharmacoepidemiological research. More specifically, the aims were:

1. To review the CPRD as a data source for pharmacoepidemiological studies of mortality
2. To conduct case studies demonstrating the use of primary care data alone for studies with a mortality outcome and when using linked data
3. To assess the impact on the study population size and follow-up time when using linked data for observational research studies
4. To evaluate the agreement between English primary care data and national death registration data in capturing vital status and the date of death

The first aim was to review the CPRD GOLD as a data source. There had been many updates and enhancements to the CPRD in recent years which had not been outlined in a concise, referenceable publication. Chapter 2.1 describes the CPRD GOLD database, the population coverage, availability of linked data and outlines the strengths and weaknesses. The key strengths are the breadth of coverage, size, long-term follow-up, representativeness and data quality. However, researchers need to be aware of the data that is not captured, such as whether prescriptions have been filled, the use of over the counter medications and treatment in secondary care. There are also certain patient groups that are missing from primary care records such as prisoners, private patients, some residential homes and the homeless.

In addition to studies in the UK, CPRD data are commonly used across the globe and particularly in the Netherlands. It is not clear whether study results based on data from this UK population would apply to the Dutch population. In chapter 2.2, the age distribution of men and women in the CPRD population were visually and numerically compared with Dutch census data. The calculation of the differences between proportional distributions of 5-year age groups showed that the CPRD population are comparable to the Dutch population in terms of age and sex up to age 75 in men and age 80 in women. This study was not able to rule out differences in ethnicity, socioeconomic status or lifestyle factors. Researchers should consider the likelihood of differences in drug exposure between the countries, which may cause variations in the populations studied.

The second aim was to conduct case studies with a mortality outcome. Chapter 3 demonstrated the use of data from CPRD to study mortality for a range of conditions. The first study examined Atrial Fibrillation (AF) patients. This condition carries an increased risk of ischaemic stroke and oral anticoagulation with warfarin can reduce this risk. However, the effectiveness of warfarin is dependent on the quality of anticoagulation with optimal risk reduction and appropriate safety occurring within the recommended International Normalised Ratio (INR) range of 2.0 to 3.0. The risks of ischaemic and intracranial haemorrhage are significantly higher for INR values outside of this range and reduced anticoagulation control in warfarin users has been associated with an increased risk of stroke and mortality.

The objective of the study in chapter 3.1 was to evaluate the association between time in therapeutic range and risk of stroke and all-cause mortality. Cox proportional hazards regression models were used to estimate the relative rates (RRs) for the associations between risk over time of stroke and mortality and time spent within therapeutic range. Mortality rates were found to be significantly lower with at least 70% of the time spent within therapeutic range. The limitations of this study were that that some patients may have used aspirin, which also has anticoagulant properties and is available over the counter in the UK. The time within therapeutic range could only be calculated for those with at least three INR measurements and some records will have been missing as not all clinics report these electronically back to GPs. The AF patients in this study were identified based on records from primary care, however since a diagnosis of AF is typically made in secondary care following review by a specialist, some patients may have been missed if the GP was not informed of a diagnosis documented at hospital.

Chapter 3.2 describes another study of mortality in atrial fibrillation patients. This study assessed newly diagnosed patients based on records from primary care and followed patients with linked hospital records, based on linked Hospital Episode Statistics (HES) data. The incidence of hospitalisation and all-cause mortality were

calculated assuming a Poisson distribution of events. The lowest rate of all-cause mortality was found during current exposure to anticoagulant treatment; higher rates were found for the periods after stopping treatment or when not treated. The same limitations in terms of aspirin use, missing INR measurements and diagnoses from primary care in chapter 3.1 also applied to this study. Additionally, this study included linked data which limited the cohort size, excluding more than half of the newly diagnosed AF patients.

Chapter 3.3. also assessed INR measurements for patients with a diagnosis of Venous thromboembolism (VTE), comprising deep venous thrombosis (DVT) and pulmonary embolism (PE). VTE is a major cause of morbidity and mortality and guidelines recommend anticoagulation immediately following diagnosis. The objectives of the study were to evaluate the pattern of anticoagulation and quality of INR control after VTE and to assess the affect these have on VTE recurrence and mortality. VTE patients were identified based on records in either primary care or hospital and Cox proportional hazards regression models were used to estimate the relative rates of hospitalisation and mortality. Overall, VTE cases were more likely to die compared to control patients without VTE. Time spent within therapeutic range was strongly associated with risk of death due to VTE and VTE recurrence, where lower risks were found in those with ≥70% time spent within therapeutic range. The strong association between levels of anticoagulation and risk of mortality was similar to that seen in chapters 3.1 and 3.2 for atrial fibrillation. The addition of diagnostic information from secondary care was a strength in this study, providing a population representative of VTE in actual clinical practice.

The fourth example of a mortality study using CPRD data examined patients treated with insulin or oral antidiabetic drugs. Concerns over the cardiovascular safety of the thiazolidinediones, particularly rosiglitazone and pioglitazone used to treat type 2 diabetes, have emerged. Chapter 3.4 evaluated the incidence of mortality, acute coronary syndrome, stroke and heart failure in cohorts of new users of a range of diabetes medications, including rosiglitazone and pioglitazone. This study found that there was a substantive heterogeneity between the populations using different classes of various diabetes medications however, comparable populations used rosiglitazone and pioglitazone. The excess risk of death over three years was calculated for rosiglitazone compared to pioglitazone, overall and by age. Higher risks for death (overall and due to cardiovascular disease) and heart failure were found for rosiglitazone compared to pioglitazone, with the risks of death highest shortly after starting treatment and disappearing after discontinuation of rosiglitazone. The outcome of death was evaluated by using both the primary care record and the linked death certificates and incidence rates were similar, however there was a loss of precision when using linked data, which had reduced the cohort size significantly.

Chapter 3 provided four examples of studies with a mortality outcome using CPRD. Linked data were included in some studies as part of the cohort definition and/or the assessment of outcomes. This may have increased the reliability of some records however had a significant impact in terms of cohort size (chapter 3.2), follow-up time (3.2) and precision (3.4) since these data are only available for a subset of the CPRD primary care population.

The third aim was to assess the impact on the study population size and follow-up time when using linked data for observational research studies. Chapter 4 looked at how using linked data can increase misclassification depending on the choice of data source(s) and how the denominator is applied. This study repeated the analyses of chapter 3.3 using various cohort definitions to explore the impact of using primary care data alone, linked hospital data and linked death registration data, as well as considering eligibility for linkage and the coverage periods for each data source. The findings showed that the choice of data source does have an impact on the results of a mortality study in VTE patients. Additionally, it may be easy to make a mistake with censoring using linked CPRD data, and this may impact on the results of a study, therefore it is important to get the denominator right to avoid misclassification and inflation of person time.

Chapter 5 focused on the agreement in the recording of deaths between data sources. There was substantial agreement between the national records from the ONS and the data from CPRD and almost half of all deaths were recorded in the HES. The comparison was based on individually linked data and it is the discordance that is of most interest. Some deaths were recorded as part of the hospital audit record or in the primary care record without being reflected in the national statistics.
The overall level of concordance was good, however given that there may be delays in a practice being informed about deaths and limited incentive to record this accurately, some consideration should be made when using primary care data alone to assess deaths. For censoring follow-up and calculating mortality rates, CPRD data are likely to be sufficient, as a delay in death recording of up to one month is unlikely to impact results significantly. Where the exact date of death or the cause are important, it may be advisable to include the individually linked death registration data from the ONS.

In conclusion, this thesis outlined the following:

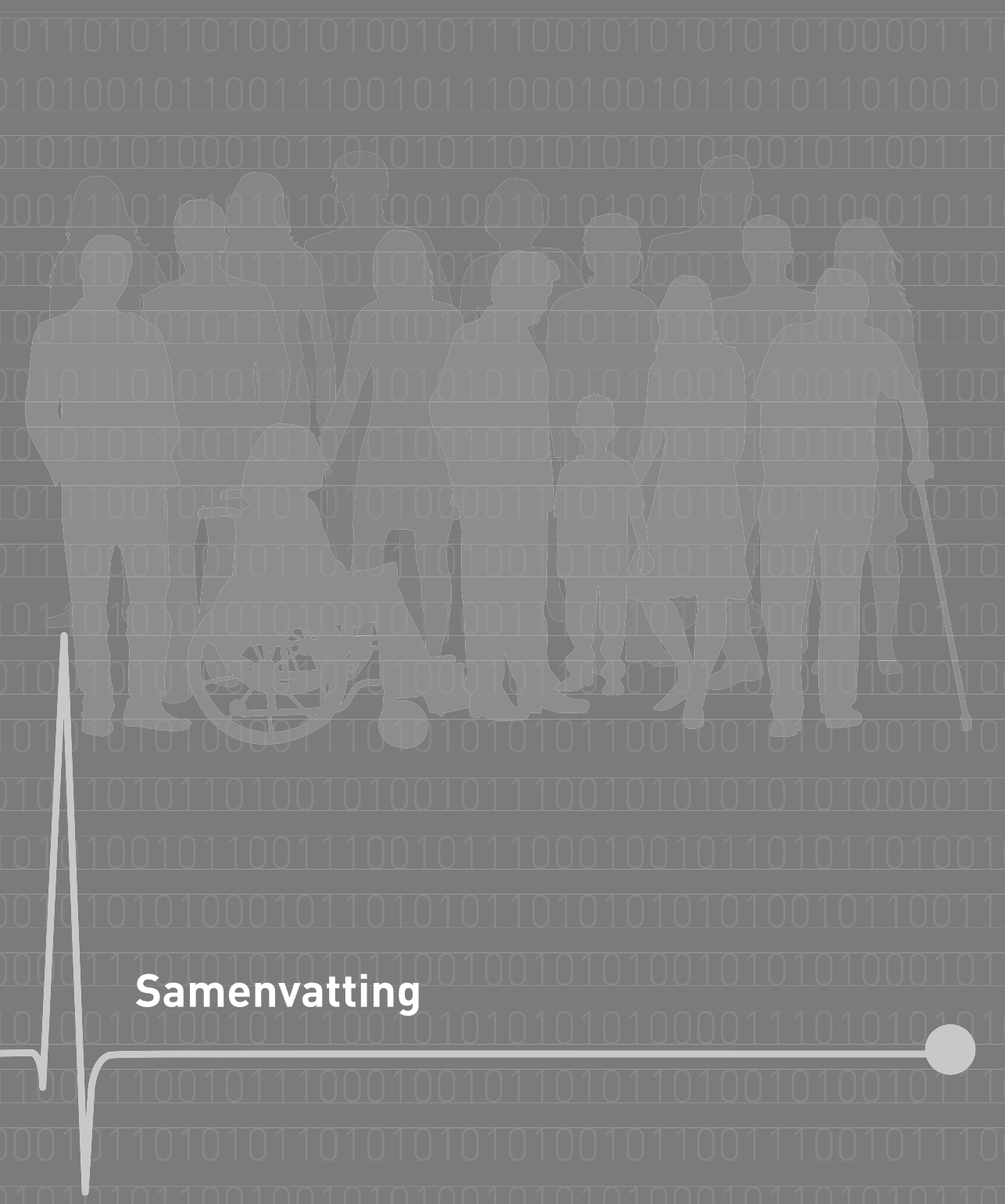
The CPRD GOLD primary care data source is representative of the UK and Dutch populations in terms of age and sex. This provides us with confidence in using this data source for mortality studies, where age and sex are important covariates. However, researchers also need to assess the external validity of this data source,

particularly for exposures or outcomes likely to be captured in secondary care or those more prevalent in populations not covered.

Linking primary care records to information from secondary care and death certificates offers the opportunity to include additional patients and provide further details on the death record. However, linked data may not be available for all subjects, therefore choosing to use linked data may have an impact on the sample size for a study as well as adding to the complexity of defining the denominator and allocating follow-up time. Researchers need to balance the opportunities and challenges when planning a study.

Information on vital status and the date of death may not be accurately recorded in primary care records. The CPRD algorithm aims to estimate a best guess for the date of death from multiple records. Concordance between the resulting death date and the date of death from the official death certificate is good, however the primary care data will often have a date of death recorded that is later than the date on the death certificate. This will have a larger impact on studies of person time.

Despite the limitations, it is important to remember that these data sources are based on documented, rather than recalled information and the availability of data linked at the individual level represents a significant revolution in the field of pharmacoepidemiology.



Samenvatting

Eén van de meest geanalyseerde uitkomsten in de farmaco-epidemiologie is overlijden. Voor volksgezondheidsonderzoek is dit vaak een 'definitieve' uitkomst, maar overlijden kan ook een belangrijk censurerende resultaat zijn in studies met een andere uitkomst. Voor observationeel onderzoek is het belangrijk om de follow-up tijd in de epidemiologische evaluatie nauwkeurig te kunnen bepalen en daarom zijn gegevens over de vitale status (dood versus levend) en de bijbehorende overlijdensdatum cruciaal.

De afgelopen 30 jaar is het gebruik van elektronische medische dossiers (EHR) voor onderzoek enorm toegenomen. Een veelgebruikte EHR-database is de Clinical Practice Research Datalink (CPRD) GOLD, gebaseerd op gegevens uit de primaire zorg van huisartsenpraktijken in het VK. Deze gegevensbron wordt vaak gebruikt door de farmaco-epidemiologie, maar er zijn beperkingen aan de gegevens omdat de gegevens voornamelijk worden verzameld voor de klinische zorg of controle. De registratie van sterfgevallen in Britse huisartsenpraktijken is dubieus, omdat er geen directe koppeling is tussen de huisartsenpraktijk en de nationale overlijdendsregister en er een beperkte motivatie kan zijn voor de correcte registratie van de vitale status van de patient en de overlijdensdatum in het patiëntendossier.

Hoofdstuk 1 schetst de beschikbare gegevensbronnen voor observatie studies van farmaco-epidemiologie in Europa, met een focus op de Britse huisartsenpraktijken in de CPRD GOLD-database. Dit is een internationaal erkende gegevensbron die wordt gebruikt voor onderzoek ten gunste van de volksgezondheid. De database met elektronische medische dossiers is gebaseerd op de Britse bevolking en is afhankelijk van de gratis toegang op het punt van zorgverlening door de National Health Service (NHS). De CPRD GOLD-database wordt vaak gebruikt voor farmaco-epidemiologie-onderzoeken, maar voordat dit proefschrift werd geschreven, zijn er geen studies geweest die de nauwkeurigheid van de status van overlijden en de geregistreerde overlijdensdatum in deze gegevensbron beoordeelden.

Het is niet duidelijk of sterfgevallen goed zijn vastgelegd in de CPRD GOLD, aangezien huisartsen mogelijk niet altijd informatie ontvangen over hun overleden patiënten. In het VK bestaat er geen direct koppeling tussen de huisartsenpraktijk en het nationale overlijdensregister, en aangezien elektronische gegevens in de primaire zorg zijn gebaseerd op informatie die oorspronkelijk is vastgelegd voor klinische zorg, kan er een beperkte motivatie zijn voor een nauwkeurige registratie van de vitale status en de overlijdensdatum. Bovendien kunnen wijzigingen in de software of in opnamepraktijken in de loop van de tijd van invloed zijn op de registratie van overlijden.

Er zijn meerdere manieren om een overlijden vast te leggen in het Vision-softwaresysteem dat wordt gebruikt door huisartsen die gegevens bijdragen aan

CPRD GOLD. Er is een algoritme ontwikkeld om geregistreerde informatie te combineren en de beste schatting voor de sterfdatum te identificeren. Dit levert een enkel afgeleide overlijdensdatum op die aan de database wordt toegevoegd. De methodologie voor het algoritme is echter niet gepubliceerd en de afgeleide datums zijn niet gevalideerd.

Ontbrekende informatie over de vitale status zal waarschijnlijk een impact hebben op de robuustheid van een onderzoek, omdat dit kan leiden tot de onjuiste toevoeging of uitsluiting van individuen in de onderzoekspopulatie en / of kan bijdragen aan een misrekening van persoonstijd. Het bijvoegen van follow-up tijd nadat een persoon is overleden, vergroot de noemer voor een studie en dit soort informatie-bias kan niet worden gecorrigeerd met analytische technieken.

Er is nu een koppeling op individueel niveau beschikbaar tussen CPRD GOLD en externe gegevensbronnen zoals de ONS-gegevens over overlijdensregistratie en de Hospital Episode Statistics (HES) geregistreerde gegevens over patiëntenzorg. Deze gegevensbronnen bieden de mogelijkheid om aanvullende informatie over diagnoses toe te voegen, extra patiënten voor een onderzoek te identificeren en overlijdensregistraties in de primaire zorg te valideren. Gekoppelddee gegevens zijn echter alleen beschikbaar voor een deel van de CPRD-populatie, op basis van huisartsenpraktijken in Engeland, die toestemming hebben gegeven. Deze beperking vermindert de potentiële cohortgrootte voor een onderzoek en voegt complexiteit toe aan het genereren van de juiste noemer.

Dit proefschrift had als doel het nut van de CPRD GOLD-database te evalueren voor studies met sterfte als uitkomst. Het algemene doel van het proefschrift was om het gebruik van CPRD-gegevens te evalueren voor het analyseren van informatie over sterfte in farmaco-epidemiologisch onderzoek. Meer specifiek waren de doelstellingen:

1. De CPRD beoordelen als een gegevensbron voor farmaco-epidemiologische studies over mortaliteit
2. Voer casestudies uit met een sterftecijfer, gebruik makend van alleen primaire zorggegevens en gebruik makend van gekoppelde gegevens.
3. Om de impact op de populatiegrootte van de studie en de follow-up tijd te beoordelen bij het gebruik van gekoppelde gegevens voor observatie onderzoek
4. Het evalueren van de overeenkomst tussen Engelse gegevens uit de primaire zorg en nationale overlijdensregistratiegegevens bij het vastleggen van de vitale status en de overlijdensdatum.

Het eerste doel was om de CPRD GOLD te beoordelen als een gegevensbron. Er waren de laatste jaren veel updates en verbeteringen van de CPRD die niet waren

beschreven in een beknopte, citeerbare publicatie. Hoofdstuk 2.1 beschrijft de CPRD GOLD-database, de dekking van de bevolking, de beschikbaarheid van gekoppelde gegevens en schetst de sterke en zwakke punten. De belangrijkste sterke punten zijn de breedte van dekking, grootte, follow-up op lange termijn, representativiteit en gegevenskwaliteit. Onderzoekers moeten zich echter bewust zijn van de gegevens die niet worden vastgelegd, zoals of recepten zijn ingevuld, het gebruik van vrij verkrijgbare medicijnen en behandeling in de secondaire zorg. Er zijn ook bepaalde patiëntengroepen die ontbreken in de registers van de primaire zorg, zoals gevangenen, particuliere patiënten, sommige verzorginstehuizen en daklozen.

Naast studies in het Verenigd Koninkrijk worden CPRD-gegevens veel gebruikt over de hele wereld en met name in Nederland. Het is niet duidelijk of onderzoeksresultaten op basis van gegevens van deze Britse bevolking van toepassing zouden zijn op de Nederlandse bevolking. In hoofdstuk 2.2 werd de leeftijdsverdeling van mannen en vrouwen in de CPRD-populatie visueel en numeriek vergeleken met Nederlandse volkstellingen. Uit de berekening van de verschillen tussen de evenredige verdeling van leeftijdsgroepen van 5 jaar bleek dat de CPRD-populatie vergelijkbaar is met de Nederlandse bevolking in termen van leeftijd en geslacht tot 75 jaar bij mannen en 80 jaar bij vrouwen. Deze studie kon verschillen in etniciteit, social-economische status of levensstijlfactoren niet uitsluiten. Onderzoekers moeten rekening houden met de waarschijnlijkheid van verschillen in geneesmiddelengebruik tussen de landen, wat variaties in de bestudeerde populaties kan veroorzaken.

Het tweede doel was om casestudies uit te voeren met een mortaliteitsresultaat. Hoofdstuk 3 demonstreerde het gebruik van gegevens uit CPRD om sterfte te bestuderen voor een aantal aandoeningen. De eerste studie onderzocht patiënten met Atrium Fibrilatie (AF). Deze aandoening heeft een verhoogd risico op ischemische beroerte en orale anticoagulatie met warfarine kan dit risico verminderen. De effectiviteit van warfarine is echter afhankelijk van de kwaliteit van de anti-stolling met optimale risico-reductie en gepaste veiligheid binnen de aanbevolen International Normalized Ratio (INR) -bereik van 2,0 tot 3,0. De risico's van ischemische en intracraniële bloeding zijn aanzienlijk hoger voor INR-waarden buiten dit bereik en verminderde anticoagulatiecontrole bij gebruikers van warfarine is in verband gebracht met een verhoogd risico op beroerte en mortaliteit.

Het doel van de studie in hoofdstuk 3.1 was om het verband tussen de tijd in de therapeutisch breedte en het risico op beroerte en mortaliteit door alle oorzaken te evalueren. Cox proportionele gevarenregressiemodellen werden gebruikt om de relatieve snelheden (RRs) te schatten voor de associaties tussen risico in de tijd van een beroerte en mortaliteit en tijd doorgebracht binnen therapeutisch breedte. Sterftecijfers bleken significant lager te zijn met ten minste 70% van de tijd

doorgebracht binnen de therapeutische breedte. De beperkingen van deze studie waren dat sommige patiënten mogelijk aspirine hebben gebruikt, dat ook antistollingseigenschappen heeft en zonder recept verkrijgbaar is in het Verenigd Koninkrijk. De tijd binnen de therapeutische breedte kon alleen worden berekend voor diegenen met ten minste drie INR-metingen en sommige gegevens ontbreken omdat niet alle klinieken deze elektronisch aan huisartsen melden. De AF-patiënten in deze studie werden geïdentificeerd op basis van huisartsengegevens, maar omdat een diagnose van AF meestal wordt gesteld na beoordeling door een specialist, kunnen sommige patiënten zijn gemist als de huisarts niet op de hoogte was van een diagnose uit het ziekenhuis.

Hoofdstuk 3.2 beschrijft een ander onderzoek naar mortaliteit bij patiënten met atriumfibrillatie. Deze studie beoordeelde nieuw gediagnosticeerde patiënten op basis van gegevens uit de primaire zorg en volgde patiënten met gekoppelde ziekenhuisgegevens op basis van gekoppelde gegevens van Hospital Episode Statistics (HES). Het aantal ziekenhuisopnames en het algemene sterfte-risico werd berekend uitgaande van een Poisson-verdeling van gebeurtenissen. Het laagste percentage algemene sterfte-risico werd gevonden tijdens behandeling met anticoagulantia; hogere percentages werden gevonden voor de perioden na het stoppen van de behandeling of wanneer niet behandeld. Dezelfde beperkingen in termen van aspirine-gebruik, ontbrekende INR-metingen en diagnoses uit de primaire zorg in hoofdstuk 3.1 waren ook van toepassing op dit onderzoek. Bovendien omvatte deze studie gekoppelde gegevens die de cohortgrootte beperkten, met uitzondering van meer dan de helft van de nieuw gediagnosticeerde AF-patiënten.

Hoofdstuk 3.3. beoordeelde tevens INR metingen voor patiënten met een diagnose van veneuze trombo-embolie (VTE), bestaande uit diepe veneuze trombose (DVT) en longembolie (PE). VTE is een belangrijke oorzaak van morbiditeit en mortaliteit en richtlijnen adviseren antistolling onmiddellijk na de diagnose. De doelstellingen van de studie waren om het patroon van antistolling en kwaliteit van INR-controle na VTE te evalueren en om het effect te beoordelen dat deze hebben op VTE-recidief en mortaliteit. VTE-patiënten werden geïdentificeerd op basis van gegevens van de primaire zorg of het ziekenhuis en Cox-modellen voor proportionele gevarenregressie werden gebruikt om de relatieve percentages van ziekenhuisopname en mortaliteit te schatten. Over het algemeen stierven VTE-gevallen eerder dan controlepatiënten zonder VTE. De tijd die binnen de therapeutische breedte werd doorgebracht, werd sterk geassocieerd met het risico op overlijden als gevolg van VTE en VTE-recidief, waarbij lagere risico's werden gevonden bij patiënten met ≥ 70% van de tijd binnen de therapeutische breedte. Het sterke verband tussen antistollingsniveaus en het risico op mortaliteit was vergelijkbaar met die in de hoofdstukken 3.1 en 3.2 voor atriumfibrillatie. De

toevoeging van diagnostische informatie uit de secondaire zorg was een sterk punt in dit onderzoek en bood een populatie die representatief was voor VTE in de klinische praktijk.

Het vierde voorbeeld van een mortaliteitsonderzoek met CPRD-gegevens onderzocht patiënten die werden behandeld met insuline of orale antidiabetica. Er is bezorgdheid ontstaan over de cardiovasculaire veiligheid van de thiazolidinedionen, met name rosiglitazon en pioglitazon die worden gebruikt om diabetes type 2 te behandelen. Hoofdstuk 3.4 evalueerde de voorkomen van mortaliteit, acuut coronair syndroom, beroerte en hartfalen bij cohorten van nieuwe gebruikers van een reeks diabetesmedicijnen, waaronder rosiglitazon en pioglitazon. Deze studie wees uit dat er een substantiële heterogeniteit was tussen de populaties die verschillende klassen van verschillende diabetesmedicijnen gebruikten, vergelijkbare populaties gebruikten echter rosiglitazon en pioglitazon. Het overmatige risico op overlijden gedurende drie jaar werd berekend voor rosiglitazon in vergelijking met pioglitazon, zowel in het algemeen als ook per leeftijd. Een hoger risico op overlijden (alle oorzaken en door hart- en vaatziekten) en hartfalen werden gevonden voor rosiglitazon in vergelijking met pioglitazon, waarbij het risico op overlijden het hoogst was kort na het begin van de behandeling en verdween na stopzetting van rosiglitazon. Het overlijden werd geëvalueerd met behulp van zowel huisartsengegevens als ook de gekoppelde overlijdenscertificaten. Incidentiecijfers waren vergelijkbaar, maar er was een verlies aan precisie in het gebruik van gekoppelde gegevens, wat de cohortgrootte aanzienlijk had verminderd.

Hoofdstuk 3 gaf vier voorbeelden van studies met een mortaliteitsresultaat met behulp van CPRD. Gekoppelde gegevens zijn in sommige onderzoeken opgenomen als onderdeel van de cohortdefinitie en / of de beoordeling van uitkomsten. Dit kan de betrouwbaarheid van sommige gegevens hebben verhoogd, maar had een aanzienlijk effect op de cohortgrootte (hoofdstuk 3.2), follow-up tijd (3.2) en precisie (3.4) omdat deze gegevens alleen beschikbaar zijn voor een subset van de CPRD primaire zorg populatie.

Het derde doel was om de impact op de populatiegrootte van de studie en de follow-up tijd te beoordelen bij het gebruik van gekoppelde gegevens voor observatie onderzoeksstudies. Hoofdstuk 4 onderzocht hoe het gebruik van gekoppelde gegevens de kans of onjuiste classificatie kan verhogen, afhankelijk van de keuze van de gegevensbron (nen) en hoe de noemer wordt toegepast. Deze studie herhaalde de analyses van hoofdstuk 3.3 met behulp van verschillende cohortdefinities om de impact van alleen het gebruik van primaire zorg gegevens, gekoppelde ziekenhuisgegevens en gekoppelde overlijdensregistraties te onderzoeken, evenals het evalueren van de wenselijkheid om de dekkingsperioden voor elke gegevensbron te koppelen. De bevindingen toonden aan dat de keuze van

de gegevensbron een impact heeft op de resultaten van een mortaliteitsstudie bij VTE-patiënten. Bovendien kunnen studieresultaten worden beïnvloed door fouten bij het gebruik van gekoppelde CPRD-gegevens, daarom is het belangrijk om een nauwkeurige noemer te krijgen en daardoor de verlenging van persoonstijd en onjuiste classificatie te voorkomen.

Hoofdstuk 5 concentreerde zich op de overeenkomst in de registratie van sterfgevallen tussen gegevensbronnen. Er was een substantiële overeenstemming tussen de nationale gegevens van de ONS en de gegevens van CPRD en bijna de helft van alle sterfgevallen werden geregistreerd in de HES. De vergelijking was gebaseerd op individueel gekoppelde gegevens en het is de discrepantie die het meest van belang is. Sommige sterfgevallen werden geregistreerd als onderdeel van het ziekenhuisverslagen of in gegevens uit de primaire zorg zonder afspiegeling in de nationale statistieken.
Over het algemeen was het overeenkomst niveau goed, maar gezien het feit dat er in een praktijk vertraging kan optreden bij het informeren over sterfgevallen en een beperkte motivatie om dit nauwkeurig vast te leggen, moet enige aandacht worden besteed aan het gebruik van alleen de gegevens uit de primaire zorg om sterfgevallen te beoordelen. Voor de censuur van de follow-up en de berekening van de sterftecijfers zijn de CPRD-gegevens waarschijnlijk voldoende, aangezien het onwaarschijnlijk is dat een vertraging in de overlijdensregistratie met maximaal een maand de resultaten aanzienlijk zal beïnvloeden. Wanneer de exacte datum van overlijden of de oorzaak belangrijk is, kan het raadzaam zijn om de individueel gekoppelde overlijdensregistratiegegevens uit het ONS op te nemen.

Tot slot schetste dit proefschrift het volgende:

De CPRD GOLD-gegevensbron voor primaire zorg is representatief voor de Britse en Nederlandse bevolking in termen van leeftijd en geslacht. Dit geeft ons vertrouwen in het gebruik van deze gegevensbron voor mortaliteitsstudies, waar leeftijd en geslacht belangrijke covariaten zijn. Onderzoekers moeten echter ook de externe validiteit van deze gegevensbron beoordelen, met name op het situaties en uitkomsten die waarschijnlijker in het ziekenhuis worden geregistreerd of die vaker voorkomen in niet-gedekte populaties.

Het koppelen van gegevens uit de primaire zorg aan informatie uit de secundaire zorg alsmede overlijdensactes biedt de mogelijkheid om extra patiënten in onderzoeken op te nemen en verdere details over het overlijdens-record te verstrekken. Het is echter mogelijk dat gekoppelde gegevens niet voor alle proefpersonen beschikbaar zijn. Daarom kan de keuze voor het gebruik van gekoppelde gegevens van invloed zijn op de steekproefomvang voor een studie, de complexiteit de noemer te definieren en de toewijzing van follow-uptijd vergroten.

Onderzoekers moeten een balans zien te vinden tussen voor- en nadelen bij het plannen van een studie.

Informatie over de vitale status en de overlijdensdatum wordt mogelijk niet nauwkeurig vastgelegd in de primaire zorg. Het CPRD-algoritme is bedoeld om een schatting te maken voor de overlijdensdatum uit meerdere gegevens. De overeenkomst tussen de primaire zorg en de gegevens van de nationale overlijdensakte in Engeland is goed, maar in de gegevens uit de primaire zorg is vaak een overlijdensdatum geregistreerd die later is dan de datum op de overlijdensakte. Dit zal een grotere impact hebben op studies van persoonstijd.

Ondanks alle hier genoemde beperkingen is het belangrijk om vast te stellen dat deze gegevensbronnen gebaseerd zijn op gedocumenteerde in plaats van op herinnerde informatie. De beschikbaarheid van op individueel niveau gekoppelde gegevens heeft voor een belangrijke revolutie in de farmaco-epidemiologie gezorgd.

Acknowledgements

Completing a PhD thesis is an exhausting, emotional struggle. You are forced to confront the doubts you have about your abilities and somehow overcome them. The mental strength and courage required often goes unrecognised and unrewarded.

However, nothing worth having comes easy.

First and foremost, I would like to express my sincere thanks and appreciation to Frank de Vries and Bert Leufkens for the excellent supervision, generous guidance and first-rate support throughout the PhD program. You never lost faith in my ability to complete this, even though I wasn't a pharmacist, I was based in the UK, and I was doing this whilst working full time. I feel privileged to have collaborated with you and want to thank you for sharing your expertise.

Secondly, I am indebted to the Clinical Practice Research Datalink for supporting me throughout and funding travel, enabling me to collaborate with key research partners and present this research at international conferences.

There are many people that I would like to thank individually, some of whom probably have no idea of their influence.

The support from the Netherlands:

Tjeerd – you started me on this track more than 10 years ago. It was a pipe dream and there was no precedent for a thesis by publication through a Dutch university, whilst working for the UK government. You were convinced it was possible and I am grateful that you did. I learned a lot about writing manuscripts from you and I was always amazed at how swiftly you could do this, even with the radio on, playing pop tunes.

Bert – you didn't give up on me. Although you are very busy, you replied my emails quickly and provided your comments and suggestions on the manuscripts rapidly and I thank you for that. This journey was long, but you made sure it was usually me holding things up. You always provided helpful feedback and valuable comments alongside warm encouragement and thoughtful guidance. We very rarely met face to face, but it was always so lovely to see you and I always left feeling buoyed with positivity.

The assessment committee: Prof. dr. O.H. Klungel, Prof. dr. M.L. Bouvy, Prof. dr. C. Neef, Prof. dr. R.H.H. Groenwold and Prof. dr. M.C.J.M. Sturkenboom – I would like to express my cordial thanks for the time you have spent reading my thesis.

The secretarial support at the department: Ineke Dinzey, Anja Elbertse and Suzanne de Visser – Thank you for refreshing my Utrecht University log in details multiple times when I continued to be a student for so much longer than most. I had so many questions about the PhD program and you were always extremely helpful.

Most importantly, Frank – I cannot thank you enough for all the hours you have put in and for the blunt truths such as, 'to be honest you should be able to do this by now'. I am grateful for the faith you bestowed in me and the support provided during the (many) times when I thought I couldn't do it. It was you who kept me on point and on message when I tended to procrastinate. I am thankful that you took on the supervising role when everything was falling apart and remain surprised that you never became exasperated with me (if you did, you never told me). You showed such great compassion when I was faced with tragedy and had to take a break from my studies, which will always be remembered. On the flip side, we had so many laughs on this journey. One of our meetings was inside the houses of parliament, during a tour in Spanish, for which neither of us understood a word. And there was of course, the legendary night at Sam's in Copenhagen. Most of all, you have been a truly great friend. This relationship will continue long after the printed thesis is finally on my bookshelf.

The support at CPRD in the UK:

Rachael – your kindness and generosity has not gone unnoticed. Thank you for your encouragement and your willingness to share all your experience on getting to the end of the PhD program. If you hadn't done it first, I probably wouldn't be where I am now.

Tim, Shivani, Jenny, Dan, Tarita, Puja, the current Observational Research team and previous colleagues Helen, Wilhelmine, Grant, Susan and Efrosini – thank you for your support on this journey, you have all contributed in some way as either co-authors, reviewers, or just by taking on some of my workload. Thank you for your encouragement and patience. It has been a pleasure to work with you all.

My co-authors:

There are too many to list individually, as over the years I have collaborated with many colleagues, from all over the world, in many different projects. However, they are all noted in the list of my published papers and it was great to work with each and every one of them. I would not have been able to accomplish this task if not for all my publications, which was most definitely a group effort.

Those who helped to make this a success:

Ralph – I am indebted to you for spending your holiday time reading my thesis and spotting errors. For many people this would have been tiresome, so I am glad that you were intrigued by some of the content.

Nanda – thank you for helping to translate my thesis summary into Dutch. Without your support there was a high chance that I would have included 'brabbeltaal'.

My paranymphs, Kay and Sarah – as soon as I read about this role, I knew who would be perfect. Thank you for being there for me throughout this journey, and especially to Utrecht, I am truly grateful. We will need more than tea and scones to celebrate after the defence.

Those who influenced me on this journey:

Willie – for introducing me to UK primary care data for research. You opened my eyes to the opportunities with this data source and I can still see such potential more than 17 years later.

Jessamy – for being an inspiration. A year or so ago I looked at your life - working full time, raising a daughter, volunteering at the weekends and studying for a PhD in your 'spare' time. If you could do this, maybe I could too. It made me dust off my paperwork and pick this up again when I had all but given up.

My family and friends:

I am especially grateful to my family and friends who have provided the sometimes much needed encouragement to enable me to complete this thesis.

Jane – for all the times you took me swimming. The quiet time in my own head while doing lengths, did wonders for clarifying sections of this thesis.

Nathalie – for all those runs around Battersea park. The break from the desk was really good for me and I always really enjoyed it when we put the world to rights.

Tracy – for always being encouraging and pushing me that bit further even when I was extremely stretched. Towards the end of this journey your faith in me really helped push me forward.

Mum – you knew more than anyone how much I wanted to do this and never stopped encouraging me to go forward.

Dad – you were there at the very start of this journey and waved me off on my first trip to Utrecht. Four days later cancer took you from us. In our very last conversation you were encouraging me to start this journey, even though you never quite understood what I was studying. Well, here we are (and there is no mention of fish).

Special thanks to my husband Brian – you have supported me throughout, with reassurance every step of the way. You were the one who encouraged me to pick up my laptop and start reading scientific papers again after the loss of our cherished baby boy. I can never thank you enough for all the meals that you hand delivered to the library and the patience you had when the writing up phase seemed to go on forever. I could not have completed this without you.

My children:

Vincent – our time together was far too short, however you still brought joy and taught me so much. Your influence will stay with me always.

Hanora – you are the rainbow after the storm, always pleased to see me with infectious laughter. You have made coming home at the end of a long day so much more exciting. There have been so many times when I have felt torn between staying in the library and spending time giggling with you. This is what I was up to when you wondered where I was.

Finally:

After 11 years, two changes of topic, two periods of maternity leave, countless hours in the library and late nights at the office, I have realised that completing a PhD thesis is the best resilience training you could ever have. I already look back and wonder how I managed to complete this thesis whilst also working full time, raising a small child and volunteering at weekends. Nobody promised it would be easy, just that it would be worth it.

The main person who thought I couldn't do this was me. I showed her.

"If you knew what you were doing it wouldn't be called research"
Einstein

List of co-authors

Co-authors of work presented in this thesis, in alphabetical order

Diana Ackermann
Boehringer Ingelheim GmbH, Ingelheim am Rhein, Germany

Dorothee B. Bartels
Boehringer Ingelheim GmbH, Ingelheim am Rhein, Germany

Krishnan Bhaskaran
London School of Hygiene & Tropical Medicine, London, UK

Andreas Clemens
Boehringer Ingelheim GmbH, Ingelheim am Rhein, Germany

Daniel Dedman
Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, London, UK

Harriet Forbes
London School of Hygiene & Tropical Medicine, London, UK

Bastian Haß
Boehringer Ingelheim GmbH, Ingelheim am Rhein, Germany

Emily Herrett
London School of Hygiene & Tropical Medicine, London, UK

Maryska L.G. Janssen-Heijnen
VieCuri Medical Centre, Venlo, The Netherlands
Maastricht University Medical Centre+, Maastricht, The Netherlands

Roy G.P.J. de Jong
VieCuri Medical Centre, Venlo, The Netherlands
Maastricht University Medical Centre+, Maastricht, The Netherlands

Hubert G.M. Leufkens
Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

Rohini Mathur
London School of Hygiene & Tropical Medicine, London, UK

Ad A.M. Masclee
Maastricht University Medical Centre+, Maastricht, The Netherlands

Tarita Murray-Thomas
Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, London, UK

Shivani Padmanabhan
Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, London, UK

Jonathan M Plumb
Boehringer Ingelheim GmbH, Ingelheim am Rhein, Germany

Nils Schoof
Boehringer Ingelheim GmbH, Ingelheim am Rhein, Germany

Suzie Seabroke
Medicines and Healthcare Products Regulatory Agency, London, UK

Efrosini Setakis
Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, London, UK

Liam Smeeth
London School of Hygiene & Tropical Medicine, London, UK

Tjeerd-Pieter van Staa
Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, London, UK
London School of Hygiene & Tropical Medicine, London, UK

Frank de Vries
Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
Maastricht University Medical Centre+, Maastricht, The Netherlands

Tim Williams
Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, London, UK

List of publications

Publications relating to this thesis are marked with an asterisk.

Quality improvement of prescribing safety: a pilot study in primary care using UK electronic health records.
Booth HP, **Gallagher AM**, Mullett D, Carty L, Padmanabhan S, Myles PR, Welburn SJ, Hoghton M, Rafi I, Valentine J.
Br J Gen Pract. 2019 Aug 29; 69(686):e605-e611.

Limitations for health research with restricted data collection from UK primary care.
Strongman H, Williams R, Meeraus W, Murray-Thomas T, Campbell J, Carty L, Dedman D, **Gallagher AM**, Oyinlola J, Kousoulis A, Valentine J.
Pharmacoepidemiol Drug Saf. 2019 Jun;28(6):777-787.

Use of Topical Tacrolimus and Topical Pimecrolimus in Four European Countries: A Multicentre Database Cohort Study.
Kuiper JG, van Herk-Sukel MPP, Castellsague J, Pottegård A, Berglind IA, Dedman D, Gutierrez L, Calingaert B, Hallas J, Sundström A, **Gallagher AM**, Kaye JA, Pardo C, Rothman KJ, Perez-Gutthann S.
Drugs Real World Outcomes. 2018 Jun; 5(2):109-116.

Cancer recording in patients with and without type 2 diabetes in the Clinical Practice Research Datalink primary care data and linked hospital admission data: a cohort study.
Williams R, van Staa TP, **Gallagher AM**, Hammad T, Leufkens HGM, de Vries F.
BMJ Open. 2018 May; 8(5):e020827.

Pharmacogenomics of statin-related myopathy: Meta-analysis of rare variants from whole-exome sequencing.
Floyd JS, Bloch KM, Brody JA, Maroteau C, Siddiqui MK, Gregory R, Carr DF, Molokhia M, Liu X, Bis JC, Ahmed A, Liu X, Hallberg P, Yue QY, Magnusson PKE, Brisson D, Wiggins KL, Morrison AC, Khoury E, McKeigue P, Stricker BH, Lapeyre-Mestre M, Heckbert SR, **Gallagher AM**, Chinoy H, Gibbs RA, Bondon-Guitton E, Tracy R, Boerwinkle E, Gaudet D, Conforti A, van Staa T, Sitlani CM, Rice KM, Maitland-van der Zee AH, Wadelius M, Morris AP, Pirmohamed M, Palmer CAN, Psaty BM, Alfirevic A; PREDICTION-ADR Consortium and EUDRAGENE.
PLoS One. 2019 Jun; 14(6):e0218115.

*The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations.
Gallagher AM, Dedman D, Padmanabhan S, Leufkens HGM, de Vries F.
Pharmacoepidemiol Drug Saf. 2019 May;28(5):563-569.

A cohort study on the risk of lymphoma and skin cancer in users of topical tacrolimus, pimecrolimus, and corticosteroids (Joint European Longitudinal Lymphoma and Skin Cancer Evaluation - JOELLE study).
Castellsague J, Kuiper JG, Pottegård A, Anveden Berglind I, Dedman D, Gutierrez L, Calingaert B, van Herk-Sukel MP, Hallas J, Sundström A, **Gallagher AM**, Kaye JA, Pardo C, Rothman KJ, Perez-Gutthann S.
Clin Epidemiol. 2018 Mar; 10:299-310.

Validity of diagnostic codes to identify hospitalizations for infections among patients treated with oral anti-diabetic drugs.
Saine ME, Gizaw M, Carbonari DM, Newcomb CW, Roy JA, Cardillo S, Esposito DB, Bhullar H, **Gallagher AM**, Strom BL, Lo Re V 3rd.
Pharmacoepidemiol Drug Saf. 2018 Oct; 27(10):1147-1150.

Postauthorization safety study of the DPP-4 inhibitor saxagliptin: a large-scale multinational family of cohort studies of five outcomes.
Lo Re V, Carbonari DM, Saine ME, Newcomb CW, Roy JA, Liu Q, Wu Q, Cardillo S, Haynes K, Kimmel SE, Reese PP, Margolis DJ, Apter AJ, Reddy KR, Hennessy S, Bhullar H, **Gallagher AM**, Esposito DB, Strom BL.
BMJ Open Diabetes Res Care. 2017 Jul; 5(1):e000400.

*****Comparability of the age and sex distribution of the UK Clinical Practice Research Datalink and the total Dutch population.
de Jong RG, **Gallagher AM**, Herrett E, Masclee AA, Janssen-Heijnen ML, de Vries F.
Pharmacoepidemiol Drug Saf. 2016 Dec; 25(12):1460-1464. doi: 10.1002/pds.4074.

Use of demographic and pharmacy data to identify patients included within both the Clinical Practice Research Datalink (CPRD) and The Health Improvement Network (THIN).
Carbonari DM, Saine ME, Newcomb CW, Blak B, Roy JA, Haynes K, Wood J, **Gallagher AM**, Bhullar H, Cardillo S, Hennessy S, Strom BL, Lo Re V 3rd.
Pharmacoepidemiol Drug Saf. 2015 Sep; 24(9):999-1003.

Determinants of saxagliptin use among patients with type 2 diabetes mellitus treated with oral anti-diabetic drugs.
Saine ME, Carbonari DM, Newcomb CW, Nezamzadeh MS, Haynes K, Roy JA, Cardillo S, Hennessy S, Holick CN, Esposito DB, **Gallagher AM**, Bhullar H, Strom BL, Lo Re V 3rd.
BMC Pharmacol Toxicol. 2015 Apr 2; 16:8.

Risk of fracture with thiazolidinediones: an individual patient data meta-analysis.
Bazelier MT, de Vries F, Vestergaard P, Herings RM, **Gallagher AM**, Leufkens HG, van Staa TP.
Front Endocrinol (Lausanne). 2013; 4:11.

Evaluation of methods to estimate missing days' supply within pharmacy data of the Clinical Practice Research Datalink (CPRD) and The Health Improvement Network (THIN).
Lum KJ, Newcomb CW, Roy JA, Carbonari DM, Saine ME, Cardillo S, Bhullar H, **Gallagher AM**, Lo Re V 3rd.
Eur J Clin Pharmacol. 2017 Jan; 73(1):115-123.

*****The Impact of the Choice of Data Source in Record Linkage Studies Estimating Mortality in Venous Thromboembolism.
Gallagher AM, Williams T, Leufkens HGM, de Vries F.
PLoS One. 2016 Feb 10; 11(2):e0148349.

*****Data Resource Profile: Clinical Practice Research Datalink (CPRD).
Herrett E, **Gallagher AM**, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L.
Int J Epidemiol. 2015 Jun; 44(3):827-36.

*****Population-based cohort study of warfarin-treated patients with atrial fibrillation: incidence of cardiovascular and bleeding outcomes.
Gallagher AM, van Staa TP, Murray-Thomas T, Schoof N, Clemens A, Ackermann D, Bartels DB.
BMJ Open. 2014 Jan 27; 4(1):e003839.

Cancer recording and mortality in the General Practice Research Database and linked cancer registries.
Boggon R, van Staa TP, Chapman M, **Gallagher AM**, Hammad TA, Richards MA.
Pharmacoepidemiol Drug Saf. 2013 Feb; 22(2):168-75.

*Quality of INR control and outcomes following venous thromboembolism.
Gallagher AM, de Vries F, Plumb JM, Haß B, Clemens A, van Staa TP.
Clin Appl Thromb Hemost. 2012 Jul; 18(4):370-8.

Risk of fracture with thiazolidinediones: disease or drugs?
Bazelier MT, Vestergaard P, **Gallagher AM**, van Staa TP, Cooper C, Leufkens HG, de Vries F. Calcif Tissue Int. 2012 Jun; 90(6):450-7.

Glucose-lowering agents and the patterns of risk for cancer: a study with the General Practice Research Database and secondary care data.
van Staa TP, Patel D, **Gallagher AM**, de Bruin ML.
Diabetologia. 2012 Mar; 55(3):654-65.

*Risks of stroke and mortality associated with suboptimal anticoagulation in atrial fibrillation patients.
Gallagher AM, Setakis E, Plumb JM, Clemens A, van Staa TP.
Thromb Haemost. 2011 Nov; 106(5):968-77.

Utilization characteristics and treatment persistence in patients prescribed low-dose buprenorphine patches in primary care in the United Kingdom: a retrospective cohort study.
Gallagher AM, Leighton-Scott J, van Staa TP.
Clin Ther. 2009 Aug; 31(8):1707-15.

Fracture outcomes related to persistence and compliance with oral bisphosphonates.
Gallagher AM, Rietbrock S, Olson M, van Staa TP.
J Bone Miner Res. 2008 Oct; 23(10):1569-75.

Resource utilization and outcomes in patients with atrial fibrillation: a case control study.
Boggon R, Lip GY, **Gallagher AM**, van Staa TP.
Appl Health Econ Health Policy. 2012 Jul; 10(4):249-59.

Use of thiazolidinediones and risk of osteoporotic fracture: disease or drugs?
Bazelier MT, **Gallagher AM**, van Staa TP, Cooper C, Leufkens HG, Vestergaard P, de Vries F.
Pharmacoepidemiol Drug Saf. 2012 May; 21(5):507-14.

*Risk of death and cardiovascular outcomes with thiazolidinediones: a study with the general practice research database and secondary care data.
Gallagher AM, Smeeth L, Seabroke S, Leufkens HG, van Staa TP.
PLoS One. 2011; 6(12):e28157.

Risk markers for both chronic fatigue and irritable bowel syndromes: a prospective case-control study in primary care.
Hamilton WT, **Gallagher AM**, Thomas JM, White PD.
Psychol Med. 2009 Nov; 39(11):1913-21.

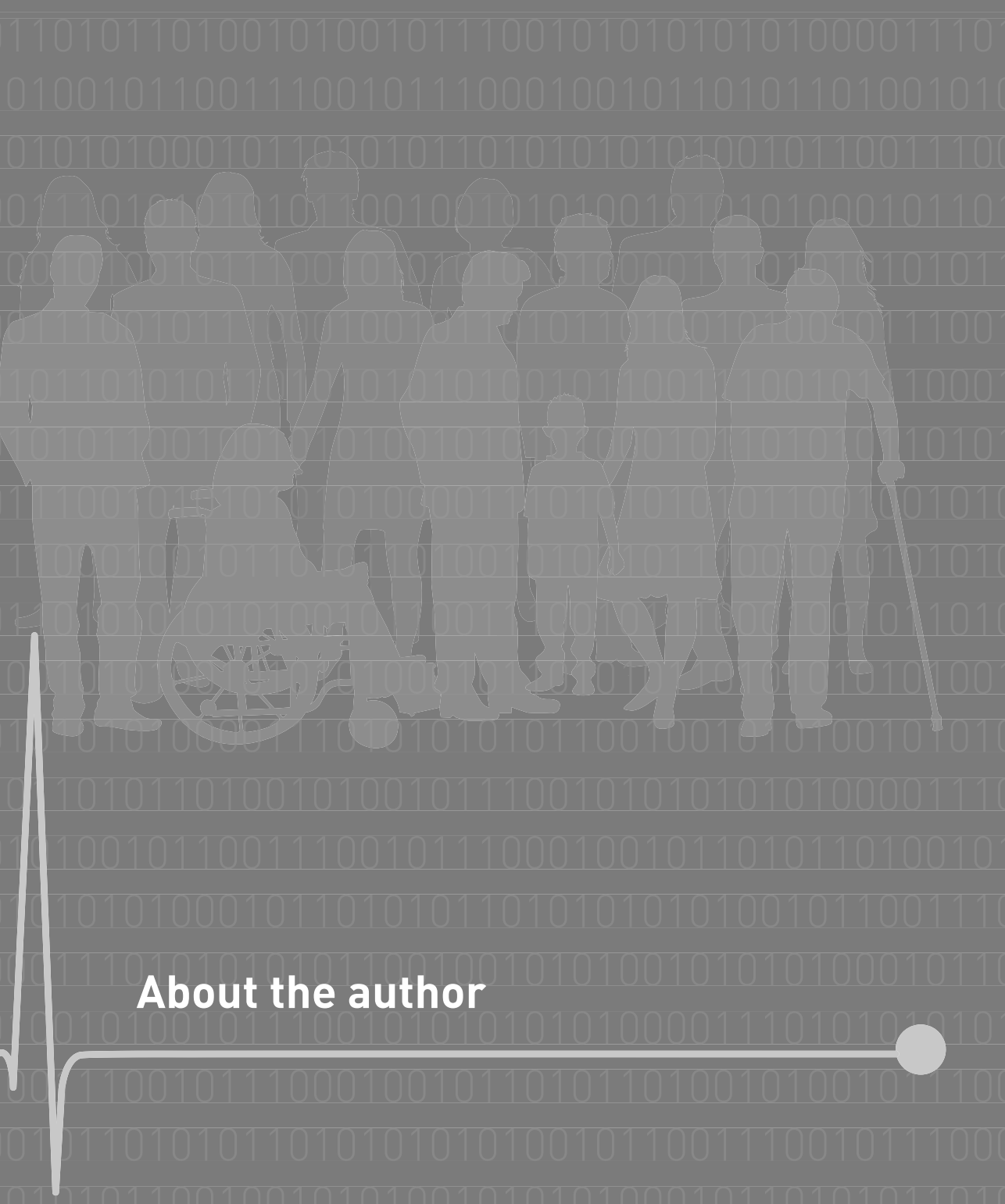
How effective are dose-adjusted warfarin and aspirin for the prevention of stroke in patients with chronic atrial fibrillation? An analysis of the UK General Practice Research Database.
Rietbrock S, Plumb JM, **Gallagher AM**, van Staa TP.
Thromb Haemost. 2009 Mar; 101(3):527-34.

Initiation and persistence of warfarin or aspirin in patients with chronic atrial fibrillation in general practice: do the appropriate patients receive stroke prophylaxis?
Gallagher AM, Rietbrock S, Plumb J, van Staa TP.
J Thromb Haemost. 2008 Sep; 6(9):1500-6.

The prognosis of different fatigue diagnostic labels: a longitudinal survey.
Hamilton WT, **Gallagher AM**, Thomas JM, White PD.
Fam Pract. 2005 Aug; 22(4):383-8.

Incidence of fatigue symptoms and diagnoses presenting in UK primary care from 1990 to 2001.
Gallagher AM, Thomas JM, Hamilton WT, White PD.
J R Soc Med. 2004 Dec; 97(12):571-5. Erratum in: J R Soc Med. 2005 Feb;98(2):88.

Is the chronic fatigue syndrome an exercise phobia? A case control study.
Gallagher AM, Coldrick AR, Hedge B, Weir WR, White PD.
J Psychosom Res. 2005 Apr; 58(4):367-73.



About the author

About the author

Arlene Marie Gallagher was born in London to Irish parents and attended a Roman Catholic primary school in Ealing, and an all-girls convent secondary school in Isleworth. She studied biological sciences at Queen Mary University of London, specialising in Zoology before completing a Masters degree in Medical Statistics at the University of Leicester. During her studies she lived as a warden in the university halls of residence, supporting undergraduate students, and spent one summer volunteering in Indonesia.

Arlene started working as a Research Statistician at the Centre for Psychiatry, Queen Mary School of Medicine and Dentistry in 2002. Whilst collaborating with Willie Hamilton (University of Bristol) she was introduced to the General Practice Research Database as a data resource. A fondness for observational data was born and she was delighted to join the GPRD (now Clinical Practice Research Datalink, CPRD) in 2004.

Arlene has collaborated on more than 30 papers using CPRD data over the last 15 years in pharmacoepidemiology, drug safety and disease epidemiology. As a Senior Researcher at CPRD, Arlene leads research studies as well as providing support to users of CPRD across the globe.

Arlene is motivated to improve public health both through maximising the potential for electronic health records (#DataSavesLives) and by encouraging individuals of every ability to physical activity and community engagement (#parkrun). As an ambassador for parkrun UK, Arlene volunteers across West London bringing new events to the family and supporting teams to deliver an initiative that aims to create a healthier and happier planet.

Arlene is married to Brian Stakelum and they have one angel (Vincent, 2017) and one rainbow (Hanora, 2018).

The research described in this thesis was completed whilst also working full time, raising a young family and volunteering in the community.