

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/173991>

**Copyright and reuse:**

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)

The contribution of primary care  
and linked data to diabetes  
pharmacoepidemiology

Helen Strongman

MSc MA (Cantab)

Thesis submitted for consideration for the degree of

Doctor of Philosophy by published work

Division of Health Sciences

Warwick Medical School

University of Warwick

United Kingdom

November 2019

## Contents

List of Illustrations and tables .....	6
Acknowledgements.....	7
Submission declaration.....	7
Statements of candidate’s contribution to the published work.....	8
Abbreviations.....	11
Summary .....	13
Background .....	15
Diabetes pharmacoepidemiology .....	15
The origins and availability of UK EHR data for research.....	16
Computerised primary care data collection in the UK.....	16
Focus on CPRD .....	17
Secondary care data and linkage to CPRD .....	18
Strengths and challenges in using primary care and linked data, with reference to diabetes pharmacoepidemiology .....	20
1) Maintaining public trust whilst facilitating health research .....	21
2) Identification of linked data study populations.....	21
3) Classification of study variables.....	21
4) Minimising confounding and channelling bias .....	24
5) Avoiding the use of future information and time-related biases .....	24
6) Comparability of data sources from different countries .....	24
Rationale for this PhD thesis and commentary linking publications .....	26
Objectives and related publications .....	27
Objective 1: Understand the methods used by NHS Digital to link CPRD primary care data to other health data sources, processing steps implemented by CPRD, and the implications of these methods for study design and reporting. Update these processes and increase transparency and awareness of these issues by publishing a peer-reviewed manuscript.....	27

Objective 2: Demonstrate general benefits, limitations and best practice in using primary care and linked health data in the field of diabetes pharmacoepidemiology. ....	28
Objective 3: Demonstrate the benefits and limitations of using linked HES APC and outpatient data to measure healthcare resource utilisation in the field of diabetes pharmacoepidemiology .....	28
Objective 4: Demonstrate the challenges, benefits and limitations of pooling linked data from different countries to assess a rare cancer outcome in the field of diabetes pharmacoepidemiology .....	28
Objective 5: Demonstrate the challenges, benefits and limitations of using linked data to assess all-cause and cause-specific mortality outcomes in the field of diabetes pharmacoepidemiology .....	29
Summary of published works .....	30
Paper 1 .....	30
Paper 2 .....	31
Paper 3 .....	32
Papers 4 and 5.....	34
Discussion.....	37
Key findings.....	37
Objective 1 .....	37
Objective 2 .....	38
Objective 3 .....	43
Objective 4 .....	44
Objective 5 .....	46
Implications for future research .....	47
Implications for policy.....	49
Conclusions .....	50
References .....	51
Appendix A: Statements of contribution signed by co-authors.....	61

Appendix B: Bibliography of works published by the candidate ..... 66

## List of Illustrations and tables

Figure 1: Stepwise prescribing of oral antihyperglycaemic drugs and insulin for  
T2DM..... 16

Figure 2: Range of linked data sources available with CPRD data ..... 20

Figure 3: Links between the strengths and challenges of using electronic health  
data described in the introduction, my objectives and published papers ..... 27

## Acknowledgements

I would like to thank my co-authors and colleagues from the Clinical Practice Research Datalink and collaborating institutions. Particular thanks go to Rachael Williams who is a co-author on 4 out of 5 submitted papers and a strong supporter of my PhD plans, and to the pan-European pioglitazone study team who never failed to entertain especially during our investigator meeting in Helsinki.

This thesis would not have been possible without the generous support of Krishnan Bhaskaran, my line manager at the London School of Hygiene & Tropical Medicine and co-supervisor for my PhD. Thanks also to my Warwick University supervisor Hema Mistry who supported me through writing this thesis.

Finally, I thank my wonderful family and friends who are always there when I need them. My Dad deserves a special mention although he is no longer with us. He believed strongly in education, would have read every word of my thesis, and would have been extremely proud!

## Submission declaration

I declare that the submitted material as a whole is not substantially the same as published or unpublished material that I have previously submitted, or am currently submitting, for a degree, diploma, or similar qualification at any university or similar institution. No parts of the works submitted have been submitted previously for any aforementioned qualification.

Word count: 10,461

## Statements of candidate's contribution to the published work

### Paper 1

Padmanabhan, S., Carty, L., Cameron, E., Ghosh, R. E., Williams, R., & **Strongman, H.** Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications.

*European Journal of Epidemiology*, 2019; 34(1): 91–99.

<https://www.ncbi.nlm.nih.gov/pubmed/30219957>

In her role as the observational research lead for linkages, Helen Strongman: assessed the potential impact of external changes to information governance and linkage processes on data quality and applicability for research and how new data sources could be used for research. She then worked with the information governance and data, tools and technology teams to update internal processes and research guidance for use of linked data. Helen proposed and promoted the writing of this manuscript, contributed substantially to the scope and structure, and made critical revisions to the content.

### Paper 2

**Strongman, H.**, D'Oca, K., Langerman, H., & Das, R. (2015). Comparison of diabetes-associated secondary healthcare utilization between alternative oral antihyperglycaemic dual therapy combinations with metformin in patients with type 2 diabetes: An observational cohort study. *Diabetes, Obesity and Metabolism*, 2015; 17(6): 573-80

*Diabetes, Obesity and Metabolism*, 2015; 17(6): 573-80

<https://www.ncbi.nlm.nih.gov/pubmed/25735201>

Helen Strongman designed the study (protocol and statistical analysis plan), developed the data management programmes, performed the statistical analysis, and drafted and revised the manuscript.

### Paper 3

Korhonen, P., Heintjes, E. M., Williams, R., Hoti, F., Christopher, S., Majak, M., Kool-Houweling, L., **Strongman, H.**, Linder, M., Dolin, P., Bahmanyar, S.

Pioglitazone use and risk of bladder cancer in patients with type 2 diabetes:

retrospective cohort study using datasets from four European countries. *BMJ (Clinical Research Ed.)*, 2016; 354: i3903.

<https://www.ncbi.nlm.nih.gov/pubmed/27530399>

This post-authorisation safety study was completed by a consortium of researchers from four European countries. Researchers in each country developed analytical datasets from the raw data and the Finnish team completed the pooled analysis. Helen Strongman programmed the UK analytical datasets using CPRD primary care and linked data sources, and ran statistical analyses as a quality assurance step for comparison with the pooled analyses. Helen also made substantial contributions to protocol design, interpretation of results and manuscript review.

#### **Paper 4**

**Strongman H.**, Korhonen P., Williams R., Bahmanyar S., Hoti F., Christopher S., Majak M., Kool-Houweling L., Linder M., Dolin P., Heintjes E.M. Pioglitazone and risk of mortality in patients with type 2 diabetes: results from a European multidatabase cohort study. *BMJ Open Diabetes Research & Care*, 2017; 5(1), e000364. <https://www.ncbi.nlm.nih.gov/pubmed/28761650>

Helen Strongman led discussions about interpretation of the findings for this analysis and took primary responsibility for drafting and revising the manuscript following comments from co-investigators and peer-reviewers. Responsibility for study design, dataset generation and statistical analysis was the same as for the primary outcome (bladder cancer, described above).

#### **Paper 5**

**Strongman H.**, Christopher S., Majak M., Williams R., Bahmanyar S., Linder M., Heintjes E.M., Bennett D., Korhonen P., Hoti F. Pioglitazone and cause-specific risk of mortality in patients with type 2 diabetes: extended analysis from a European multidatabase cohort study. *BMJ Open Diabetes Research & Care*, 2018; 6(1):e000481. <https://www.ncbi.nlm.nih.gov/pubmed/29379607>



Helen Strongman led discussions about interpretation of the findings for this analysis and took primary responsibility for drafting and revising the manuscript following comments from co-investigators and peer-reviewers. Responsibility for study design, dataset generation and statistical analysis was the same as for the primary outcome (bladder cancer, described above).

**Copies of these statements of contribution, signed by co-authors can be found in Appendix A.**

## Abbreviations

A&E	Accident & Emergency
APC	Admitted Patient Care
BMI	Body Mass Index
CPRD	Clinical Practice Research Datalink
EHR	Electronic Health Records
EMA	European Medicines Agency
EMIS	Egton Medical Information Systems
GP	General practitioner
HbA1c	Hemoglobin A1c
HES	Hospital Episode Statistics
HR	Hazard Ratio
HTA	Health Technology Assessment
ICD	International Classification of Diseases
MHRA	Medicines and Healthcare products Regulatory Agency
MPR	Medication Possession Ratio
NCRAS	National Cancer Registration and Analysis Service
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
OHA	other oral antihyperglycaemic agent
ONS	Office of National Statistics
OPCS	Office of Population Censuses and Surveys Classification of Interventions and Procedures codes
PASS	Post-authorisation Safety Study
PROactive	prospective pioglitazone clinical trial in macrovascular events
QOF	Quality and Outcomes Framework
RCT	Randomised Control Trial
SU	Sulphonylurea
SACT	Systemic Anti-Cancer Treatment data
T1DM	Type 1 diabetes mellitus
T2DM	Type 2 diabetes mellitus
THIN	The Health Improvement Network
WHO	World Health Organisation



## Summary

Pharmacoepidemiology is the study of the use, effectiveness and safety of drugs in large populations. Studies commonly use data that are routinely collected in clinical care including electronic health records (EHR) from general practice linked to data collected in hospitals and other settings[1]. The strengths and challenges of using these data have often been demonstrated in studies of diabetes[2,3], a prevalent chronic health condition characterised by hyperglycaemia leading over time to microvascular and macrovascular disease[4].

The Clinical Practice Research Datalink (CPRD) is one of the main research data services worldwide providing de-identified primary care data linked to other health data sources for health research[5]. The publications that form this thesis were completed during my role in the Observational Research team at CPRD from 2011 to 2017. My aim was to improve the technical and information governance aspects of linking primary care data to other health datasets in the UK, inform development of new linked data sources, and demonstrate the value and best practice use of primary care and linked data through applied diabetes pharmacoepidemiology studies.

Paper 1[6] describes the methods used by National Health Service (NHS) Digital to link CPRD primary care data to other health data sources, the processing steps implemented by CPRD, and the implications of these methods for study design and reporting. I was the lead observational researcher guiding updates to these processes including changes to information governance processes and the addition of new datasets. Key messages resulting from this work are that CPRD and NHS Digital operate data linkages under a well governed and robust framework. These linkages enable a broader range of pharmacoepidemiology research, improved variable definitions, and obviate the need to link data for individual studies. Metadata are supplied to inform applied research design including selection of denominator populations and study periods. Further research is required to compare alternative linkage methodologies and explore potential biases introduced through the linkage process.

My applied research includes a study comparing secondary healthcare resource utilisation in patients prescribed alternative second line type 2 diabetes (T2DM)

regimens (Paper 2[7]) and a post-authorisation safety study (PASS) investigating bladder cancer and mortality outcomes following pioglitazone prescribing (Papers 3 to 5[8–10]). These examples demonstrate that primary care data can be used to identify patients with diabetes and a wide range of related exposures, outcomes and covariates for epidemiological research. Identification of conditions that are also treated in secondary care, secondary care resource utilisation and cause-specific mortality can be improved with the use of linked data. Multiple decisions and assumptions are required to select data sources and study populations, define study variables and apply statistical analysis methodologies that account for missing data, avoid time-related biases, and minimise confounding. Methodological research is available to guide some of these decisions but should be considered in the context of the individual study and extended or updated if insufficient evidence exists. Applying the same methodology to pooled linked data sources from multiple countries can increase precision in investigating rare outcomes but does not eliminate heterogeneity due to systematic differences in diabetes patients and treatments between countries, and differences in data recording.

In summary, the work presented in this thesis shows primary care and linked health data to be important resources in diabetes pharmacoepidemiology, with standard linkages adding value to the data. My contribution to the establishment of linkage and information governance processes is described. My applied research addressed key clinical questions, and demonstrated the importance of developing and following best practice to optimise scientific quality, and increase confidence in these resources among the general public and policy makers.

## Background

### Diabetes pharmacoepidemiology

During development, pharmaceutical interventions are rigorously tested in a series of pre-clinical and clinical studies that culminate in large scale randomised control trials (RCTs). These provide strong evidence of the short- to medium- term efficacy of drugs under controlled conditions in relatively homogeneous populations. These studies are the lynch pin of marketing authorisation applications. However, regulators and Health Technology Assessment (HTA) associations increasingly require further research pre- and post- authorisation demonstrating the safe and cost-effective use of pharmaceutical interventions in real world clinical care[11]. These studies come under the umbrella of pharmacoepidemiology and commonly use secondary data sources such as EHRs which are collected during routine clinical care.

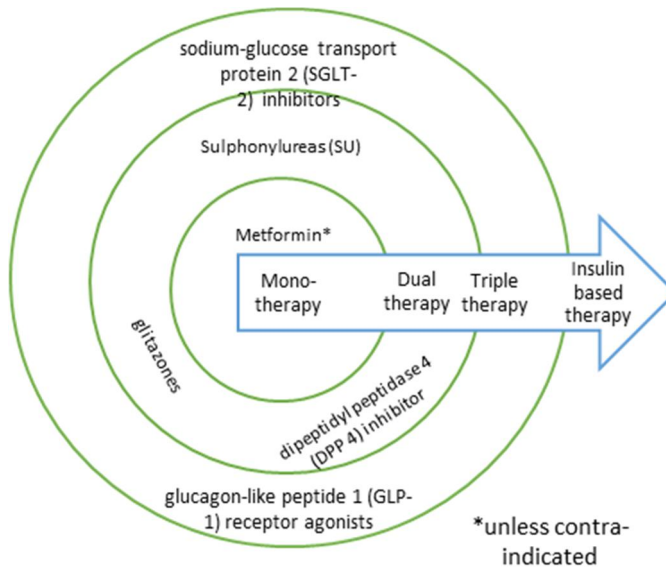
The strengths and challenges of EHR research have often been demonstrated in the field of diabetes pharmacoepidemiology[2,3] which has been one of the main topics of pharmacoepidemiological research using UK electronic healthcare data[5].

Diabetes is a chronic health condition characterised by hyperglycaemia which leads, over time, to microvascular and macrovascular disease; it is a leading and growing cause of morbidity and mortality worldwide and has a substantial socio-economic impact on individuals and communities[4]. Because of this, the United National General Assembly declared diabetes as an international public health issue in 2006[12].

T2DM is the most common form of diabetes. Risk factors including aging, ethnicity, genetics, excess adiposity, diet, sedentary behaviour and smoking lead to ineffective use of insulin by the body (insulin resistance) and resultant hyperglycaemia. Prevention is possible through the identification and monitoring of high risk groups together with interventions that encourage healthy diets, exercise and smoking cessation. Once diagnosed, T2DM is treated by diet and exercise, where possible, followed by stepwise prescribing of oral antihyperglycaemic drugs and insulin at later stages (Figure 1) [13]. Type 1 diabetes (T1DM) is the result of insufficient production of insulin by the pancreas,

the cause of which is unknown. Onset most commonly occurs in childhood and is not thought to be preventable. In gestational diabetes, hyperglycaemia is temporarily raised to a level below the cut-off for diabetes[4]. Insulin is the mainstay of treatment for T1DM[14].

Figure 1: Stepwise prescribing of oral antihyperglycaemic drugs and insulin for T2DM



## The origins and availability of UK EHR data for research

### Computerised primary care data collection in the UK

The value of UK electronic health research data derives from the creation of the NHS providing free healthcare at the point of delivery in 1948 and pioneers who developed the use of General Practice computing systems. Counts of registered patients[15] exceed Office of National Statistics (ONS) mid-year population estimates[16]. The vast majority of people living in the UK are therefore likely to be registered with a general practice. An estimated 98-99% of secondary care is also funded by the NHS[17]. Routinely collected UK data from a single system can therefore be used to conduct research that is representative of the general population.

Computers were first used in general practice in 1970 and comprehensive coding systems developed in the early 1980s[18]. By 1996, 96% of general practices were computerised[18] and a unique identifier, the NHS number was issued to all registered patients [19]. Today, General Practitioners (GPs) in the UK choose between interoperable software systems approved by NHS Digital in England and devolved bodies in Scotland, Northern Ireland and Wales. The main systems are

The Phoenix Partnership SystemOne, Egton Medical Information Systems (EMIS) Web, Vision, and Microtest Evolution[20]. General practice staff enter administrative and clinical data to manage patient care using a combination of coded information, free text, and numerical values (e.g. Body Mass Index (BMI), blood glucose and cholesterol). In addition to symptoms, diagnoses and immunisations recorded during consultations in general practice, data entered includes:

- diagnoses made in hospital and communicated back to the GP mostly via written communication, where this is considered to affect the ongoing clinical care of the patient;
- administrative information about referrals to other healthcare providers;
- test data electronically transmitted between laboratories and general practices;
- electronic prescription records issued in primary care;

Until recently, all systems used Read clinical codes[21] to identify diagnoses, care events, tests, clinical observations and lifestyle information. A new international clinical coding system, SNOMED CT, has more recently been mandated and will dominate in the near future[22].

The main UK primary care databases are the CPRD, The Health Improvement Network (THIN)[23] and QResearch[24]. Between 2004-2013, 1,296 scientific papers were published using these databases, 63.6% of which were based on CPRD data. Output from CPRD is increasing at a faster rate than THIN and QResearch[25]. These data are used by universities, governments and pharmaceutical companies worldwide to support a wide range of research[5].

#### Focus on CPRD

CPRD is a not for profit government research service, housed within the Medicines and Healthcare products Regulatory Agency (MHRA) and additionally sponsored by the National Institute for Health Research ([www.cprd.com](http://www.cprd.com)). CPRD gains annual ethics approval from the UK's Health Research Authority Research Ethics Committee permitting the receipt and supply of data for public health research. Researchers using the data for observational research must obtain approval from the MHRA's Independent Scientific Advisory Committee (ISAC) for database research for individual study protocols and observe contractual CPRD



data governance requirements. Research user license fees allow CPRD to recoup the cost of collecting and delivering data[26].

CPRD has provided de-identified health data for research for more than 30 years. Until October 2017, CPRD supplied data from a single primary care database, CPRD GOLD[27]. This includes data collected from practices using Vision software. In October 2017, CPRD launched a second database, CPRD Aurum based on data collected with EMIS software[26].

CPRD GOLD includes data from participating general practices in England, Wales, Scotland and Northern Ireland who have contributed data between 1987 and the present. Data are structured in the following files: patient (gender, year of birth, patient registration dates), consultation, clinical, additional clinical detail, referral, immunisation, test and prescription. Personal identifiable data and information entered into the patient data as free text are not collected, with the exception of dosing information, which are anonymised by CPRD. CPRD derive data quality markers including a patient level flag marking permanently enrolled patients with complete and consistent registration data, and a practice level up to standard date estimating the start of continuous data collection. In January 2014, CPRD GOLD held data from 674 practices covering 79 million person-years of follow-up; median (IQR) follow-up was 5.1 years (1.8-11.1) for individual patients. The database is broadly representative of the UK population in terms of age, sex and ethnicity [27].

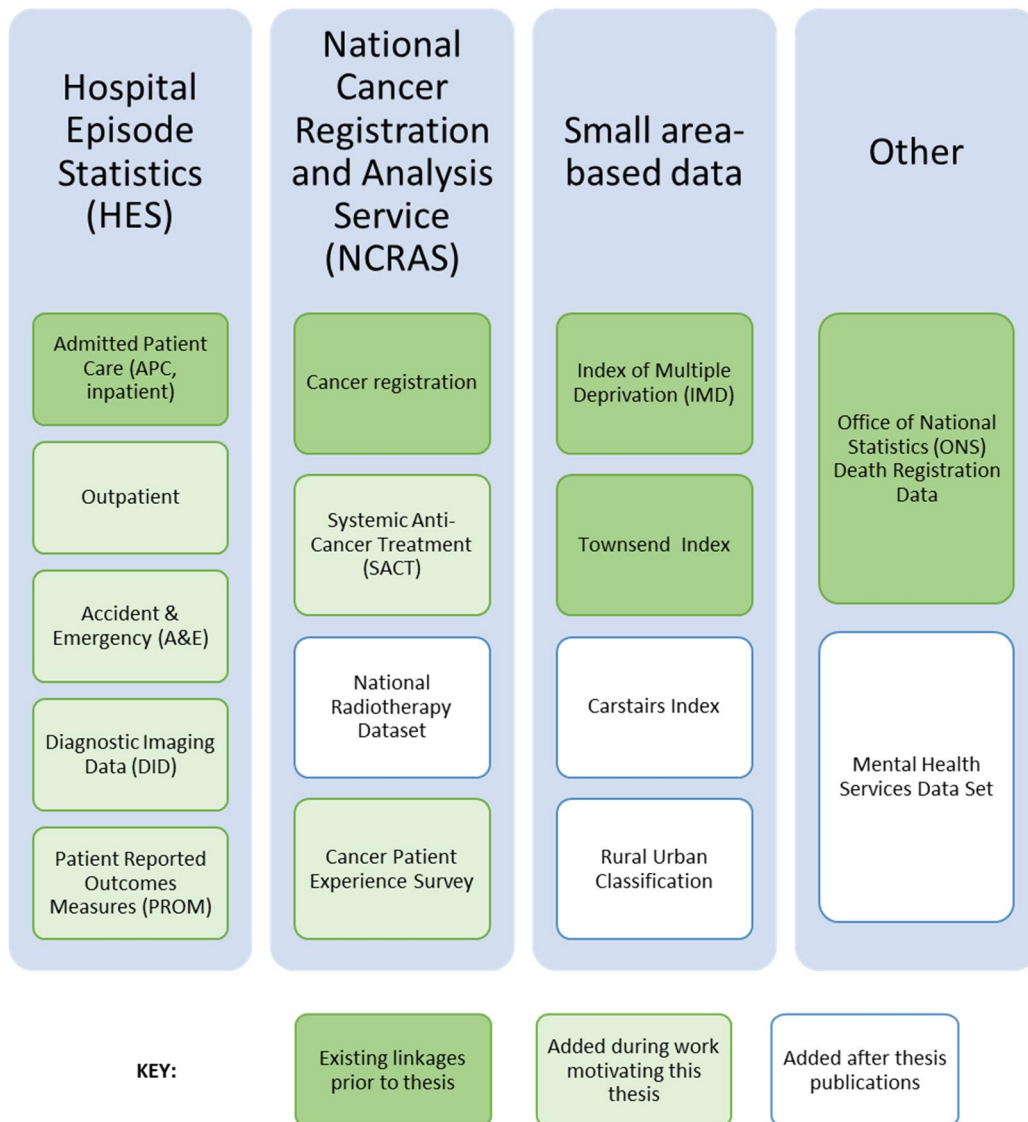
#### [Secondary care data and linkage to CPRD](#)

Hospital Episode Statistics data (HES) describe hospital activity in England and are also used to determine how much hospitals are reimbursed for care[28]. Clinical coders enter clinical and administrative data describing Admitted Patient Care (APC), Outpatient visits and Accident & Emergency (A&E) visits into hospital electronic patient information databases using information from discharge summaries. Data are transferred to a national warehouse within NHS Digital. Diagnostic and procedural data are submitted using International Classification of Diseases (ICD)-10 and Office of Population Censuses and Surveys Classification of Interventions and Procedures codes (OPCS). Additional HES datasets include diagnostic imaging and patient reported outcome measures for specific conditions.

National disease registries are also available including the National Cancer Registration and Analysis Service (NCRAS), which is housed by Public Health England[29]. Historically, these data were restricted to cancer registration and tumour details. More recently, chemotherapy, radiotherapy and cancer experience surveys have been added.

CPRD data are regularly linked to the above sources of health data. Area based deprivation data, such as Index of Multiple Deprivation quintiles, are also available through linkage to patient and practice postcodes. These data can be used as a proxy for socio-economic status. Data are linked through a trusted third party to maintain separation of personal identifiers from de-identified research data, and are only available for patients with valid personal identifiers registered in English practices that have consented to linkage [6]. All CPRD Aurum practices meet these criteria. The range of linked data sources available are described in Figure 2.

Figure 2: Range of linked data sources available with CPRD data



Strengths and challenges in using primary care and linked data, with reference to diabetes pharmacoepidemiology

The major strength of EHR data research lies in the real time collection of clinical data for large samples of patients, which for countries with national healthcare systems is representative of the general population. This increases the power of research allowing the study of rare diseases, drug regimens and outcomes, and the influence of a wide range of patient characteristics on drug safety and efficacy.

### 1) Maintaining public trust whilst facilitating health research

Use of anonymised patient data for research relies on trustworthy collection, dissemination and use of data. It is essential to uphold robust information governance practices while encouraging research that benefits public health.

### 2) Identification of linked data study populations

CPRD GOLD is an open cohort with follow-up determined by dates when individual patients enrol in and leave practices, and from which practices start to record continuous data (CPRD estimated Up-to-Standard date) and stop contributing to CPRD. This is further complicated by the use of linked data, which are only available over a specified data coverage period and for patients who were included in the linkage process. The research impact of this, and the linkage process itself, should be carefully considered by data providers so that appropriate guidance can be provided to researchers.

### 3) Classification of study variables

Use of routinely collected data for clinical care creates challenges in deriving study variables such as medical conditions, treatment patterns, additional covariates, healthcare resource utilisation and mortality. Variable definitions piece together multiple coded records associated with a pseudonymised patient identifier and a date. Codes may have been recorded at different times during the patient's disease pathway and be of lesser or greater specificity. Resultant misclassification may lead to information and selection bias, especially where the level of misclassification differs between comparison groups.

#### *3a) Identification of patients with diabetes*

CPRD GOLD includes diagnostic Read codes for type 1, type 2, gestational and unspecified diabetes; prescription records for oral anti-hyperglycaemic drugs and insulin; Read codes describing the care pathway and complications (e.g. diabetes monitoring, diabetic retinopathy), glucose test results and prescriptions for personal glucose monitoring devices. Algorithms have been published using different combinations of these codes to identify patients with diabetes and to differentiate between T1DM and T2DM[30–33]. Use of Read codes varies widely between practices and over time; different algorithms therefore lead to different observed patterns of incidence over time, and potentially to biased estimates of the association of diabetes with exposures and outcomes[33]. Incentivisation

through the Quality and Outcomes Framework (QOF) [34] has improved data availability since recording of diabetes was required in 2004 and recording of diabetes type was mandated in 2006.

### *3b) Identification of patients with other medical conditions*

External validation studies have estimated high positive predictive values in several disease areas suggesting a low false positive rate[35]. GPs are typically asked to complete questionnaires confirming diagnoses identified in the database. In the past, it was also possible to request anonymised copies of paper medical records from participating general practices or anonymisation of free text data for confirmation; this is no longer possible due to changes in information governance policies. Sensitivity cannot be measured using these techniques.

Where linked data are available, concordance studies can compare recording of diseases in different sources; studies of this type have demonstrated the benefits of using linked data to identify conditions such as myocardial infarction and gastrointestinal bleeding[36,37] that are treated in inpatient care and may not be entered as a coded record especially if they are not considered to affect the ongoing care of the patient in primary care or the patient died in hospital. Completeness of recording of cancers, especially those that require follow up in the community, is much greater in primary care[38]. Here, the value of linked data lies more in specific coding of the cancer site and the availability of information about stage of disease which is rarely coded in primary care data.

### *3c) Developing treatment algorithms*

Detailed information is available for each prescription issued by general practices. Separate records for each drug include a drug substance code, prescription date, quantity to be dispensed by the pharmacy and dose to be taken by the patient (e.g. take one tablet daily). Prescriptions may be issued automatically over regular time periods or individually at the patient's request. Prescriptions often overlap or are issued with gaps greater than the prescribing duration due to factors such as changes of dose during the prescription window, management of multiple drugs, adherence and patient holidays[39]. Researchers develop algorithms to estimate duration of individual prescriptions, duration of continuous prescribing and changes to drug regimens over time using the information available. These complex algorithms are rarely published in detail within applied research

protocols or publications and there is no standard framework[40]. Elucidation of treatment patterns is particularly complicated in T2DM pharmacoepidemiology due to the stepwise nature of treatment decisions. Dispensing information is not available with CPRD data and there is no record of whether or not the patient took the drug. Measures of adherence are therefore limited to indirect methods such as the Medication Possession Ratio (MPR) which estimates the proportion of time covered by prescriptions[41].

### *3d) Measuring additional covariates and accounting for missing data*

Lifestyle, test, and observational variables are important predictors of T2DM and diabetes progression. Validation studies have demonstrated the strengths of CPRD in measuring smoking status and BMI, with improvements in recording due to QOF which has also increased testing of Haemoglobin A1c (HbA1c)[42,43]. Recording of these variables is not complete and missing data must be accounted for through methods such as complete case analysis or multiple imputation[44,45]. Multiple imputation involves using the observed data to create several datasets with plausible predicted values in place of missing data, reflecting the assumed distribution from which the missing data came. Statistical analysis is then completed in each dataset and combined using Rubin's rules to take account of uncertainty in imputed values[46]. The validity of this method relies on the assumption that systematic differences between patients with missing data and complete values can be fully explained by the observed data i.e. missing at random (MAR). Other covariates such as diet, exercise are not commonly recorded in primary care.

### *3e) Using linked data to measure healthcare resource utilisation*

The availability of healthcare resource utilization data has been much improved by the addition of linked datasets[47]. Prior to starting the work summarised in this thesis, the benefits and limitations of using HES outpatient data to capture resource utilization had not been demonstrated.

### *3f) Using linked data to measure all cause and cause-specific mortality*

Research has shown the marginal benefit of using ONS mortality data to measure date of death[48]. There are no guidelines recommending best practice use of the cause-specific death data.

#### 4) Minimising confounding and channelling bias

Randomisation of patients to intervention groups in well designed and conducted RCTs ensures that exposure is not associated with baseline patient characteristics that influence the outcome under study[49]. In real world clinical practice, choice of pharmaceutical interventions by clinicians and patients is strongly influenced by patient characteristics. Observed associations between pharmaceutical interventions and outcomes may therefore be caused by these differences rather than causal relationships. EHR study methodologies must identify and account for confounding[50]. For example, the use of stepwise treatment algorithms in T2DM causes channelling, a type of confounding by indication. Metformin, the recommended first line drug, is systematically prescribed to patients at earlier stages of diabetes and less risk of a wide range of outcomes including cancer, than alternative anti-diabetic drugs such as pioglitazone. A naïve comparison of these drugs would falsely suggest that metformin has a protective effect against cancer compared to alternative anti-diabetic drugs[3]. Traditionally, epidemiologists have used a combination of matching of comparison groups and adjustment in multivariable regression models to minimise confounding in observational studies. In recent years, propensity score approaches have become a popular method in EHR research[51]. The propensity score is defined as the probability of being exposed to the treatment of interest given everything that is known about the individual[52].

#### 5) Avoiding the use of future information and time-related biases

Biases may also be introduced in EHR studies through the use of future information or misclassification of exposure time as described by Farmer et al. and Suissa and Azoulay[2,3]. These include immortal time bias where exposed status is assigned after the start of follow-up. Patients with an outcome between start of follow-up and exposure are by definition counted as unexposed, whereas equivalent patients without an early outcome are classified as exposed; this typically leads to spurious protective effects being observed.

#### 6) Comparability of data sources from different countries

It is important to replicate studies in different countries and healthcare settings to understand whether findings are generalisable to the global population. Substantial variation in the purpose and tools used for data collection creates

challenges in reliably replicating study designs and understanding whether differences are due to populations differences or the databases used[53].

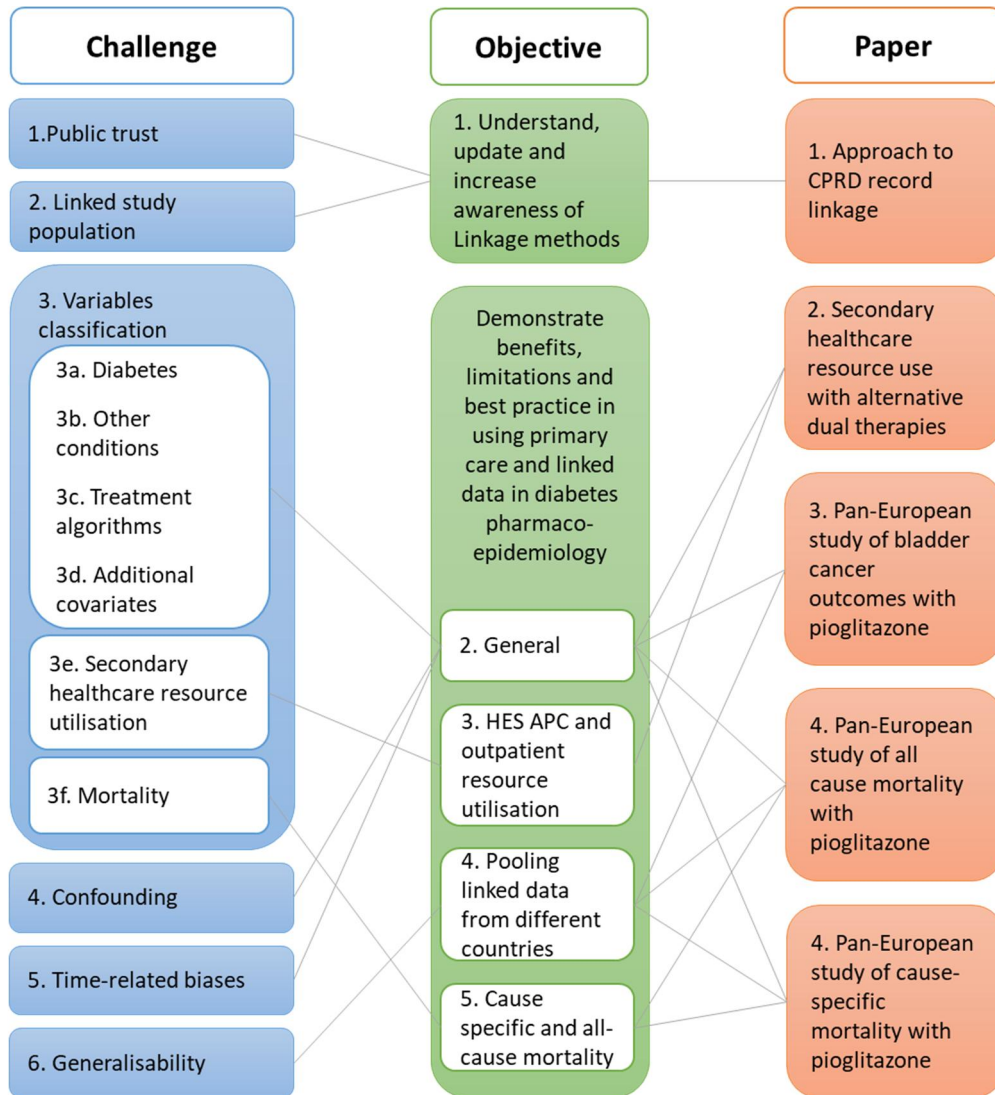


## Rationale for this PhD thesis and commentary linking publications

The research described in this thesis was completed as part of my role in the Observational Research team at CPRD. In this role, I focussed on improving, extending and demonstrating the research benefit of linking primary care data with other health databases. I conducted applied examples in the field of diabetes pharmacoepidemiology, in collaboration with industry researchers.

Based on this work, this thesis aims to improve understanding of the technical and information governance aspects of linking primary care data to other health datasets in the UK, inform development of new linked data sources, and demonstrate the value and best practice use of primary care and linked data in diabetes pharmacoepidemiology. Links between the strengths and challenges of using electronic health data described in the introduction, my objectives and published papers are described in Figure 3.

Figure 3: Links between the strengths and challenges of using electronic health data described in the introduction, my objectives and published papers



### Objectives and related publications

Objective 1: Understand the methods used by NHS Digital to link CPRD primary care data to other health data sources, processing steps implemented by CPRD, and the implications of these methods for study design and reporting. Update these processes and increase transparency and awareness of these issues by publishing a peer-reviewed manuscript. I studied the process through which CPRD primary care data are linked to other health data resources and assessed the potential impact of external changes to information governance and linkage processes on data quality and denominator populations. I then worked with the data, tools and technology and information governance teams to update CPRD's internal processes. The full process is

described in [Paper 1](#) together with implications for research including recommendations regarding the definition of denominator populations in linked data studies.

Objective 2: Demonstrate general benefits, limitations and best practice in using primary care and linked health data in the field of diabetes pharmacoepidemiology.

I conducted applied studies in two areas of diabetes pharmacoepidemiology: healthcare resource utilisation ([Paper 2](#)) and drug safety ([Papers 3 to 5](#)). I will use all four papers to discuss general benefits, limitations and best practice use of primary care and linked health data in the field of diabetes pharmacoepidemiology with reference to the following challenges: identifying patients with T2DM, identifying patients with other conditions, describing treatment algorithms, measuring lifestyle covariates and accounting for missing data, minimising confounding and channelling bias, and avoiding the use of future information and time-related biases.

Objective 3: Demonstrate the benefits and limitations of using linked HES APC and outpatient data to measure healthcare resource utilisation in the field of diabetes pharmacoepidemiology

[Paper 2](#) compares diabetes-associated secondary healthcare utilization in patients with T2DM prescribed alternative oral dual therapy combinations. This was the first applied study to use CPRD linked HES outpatient data. I will demonstrate the challenges, benefits and limitations of using HES APC and outpatient data to measure healthcare resource utilisation outcomes.

Objective 4: Demonstrate the challenges, benefits and limitations of pooling linked data from different countries to assess a rare cancer outcome in the field of diabetes pharmacoepidemiology

[Papers 3, 4 and 5](#) report findings from a pan-European PASS investigating bladder cancer and mortality risk with pioglitazone use in T2DM. I will discuss the related challenges, benefits and limitations with reference to heterogeneity in the results between countries and between linked and unlinked UK datasets.

Objective 5: Demonstrate the challenges, benefits and limitations of using linked data to assess all-cause and cause-specific mortality outcomes in the field of diabetes pharmacoepidemiology

I will discuss the challenges, limitations and benefits of studying mortality outcomes using linked and unlinked data (Papers 4 and 5).

## Summary of published works

### Paper 1

Approach to record linkage of primary care data from CPRD to other health-related patient data: overview and implications (Padmanabhan et al, 2019[6])

#### *Introduction*

CPRD's routine linkage of primary care data to secondary health data sources extends the scope and quality of CPRD studies. The linkage process involves a trusted third party, NHS Digital, and further internal data processes which uphold essential principles regarding the safe use of health data and inform robust study design. These processes were updated recently and have not previously been described in detail in a peer-reviewed paper.

#### *Objective*

To update and describe robust and transparent linkage processes for CPRD data that promote the use of best practice methodologies for applied study design, and to encourage researchers to consider the impact of the linkage process on their research.

#### *Methods*

Paper 1 briefly describes CPRD data governance and ethics requirements including the general practice 'opt-in' and patient 'opt-out' system; ethics, legal and data sharing approvals obtained by CPRD and data access conditions for researchers.

The data flow between General Practices, external data custodians, NHS Digital and CPRD is summarised highlighting:

- separation of personal identifiers from clinical data;
- the eight-step deterministic linkage method;
- linked data sets available at the time of publication;
- processes, file structures and metadata developed by NHS Digital and CPRD to support high quality research and the identification of denominator populations.

Data governance and linked data flow is under constant review by CPRD and partner organisations and require updating from time to time. Motivations for updates that I was involved in included:

- Redevelopment of the linkage process by NHS Digital in 2014
- The incorporation of Read codes for dissent associated with the care.data project in 2013
- Redevelopment of CPRD data processing including a comprehensive quality assurance process
- Addition of new linked datasets including HES Outpatient, Systemic Anti-Cancer Treatment data, the Cancer Patient Experience Survey and recent versions of the Index of Multiple Deprivation

### *Conclusions*

This paper increases transparency and awareness of the linkage process. This, together with metadata developed through the linkage process, can be used to inform best practice study design and understanding of the strengths and limitations of using linked CPRD data sources. The manuscript can also be used by researchers planning novel health data linkages internationally.

### *Paper 2*

Comparison of diabetes-associated secondary healthcare utilization between alternative oral antihyperglycaemic dual therapy combinations with metformin in patients with type 2 diabetes: An observational cohort study (Strongman et al, 2015[7])

### *Introduction*

When this research was conducted, T2DM treatment guidelines recommended the addition of sulphonylureas (SUs) to first line metformin monotherapy when glycaemic control was inadequate[13]. Newer, alternative add-ons were recommended when SUs were not tolerated or contradicted (e.g. due to risk of hypoglycaemia). SUs have also been associated with an increased risk of long-term adverse outcomes including cardiovascular disease [54–58]. These associations may increase hospital attendance by patients prescribed SUs potentially offsetting higher prescribing costs with the newer agents.

### *Objective*

To compare diabetes-associated secondary healthcare utilization in patients with T2DM prescribed sulphonylureas versus other oral antihyperglycaemic agents (OHAs) as an add-on to metformin monotherapy.

### *Methods*

I conducted a propensity score matched cohort study in adults with T2DM using CPRD GOLD and HES (APC and outpatient) data. The exposed and control groups were initiated on SU or an alternative OHA respectively after first line metformin monotherapy during the study period (April 2003-March 2012). The primary outcome was diabetes-associated secondary healthcare admissions and outpatient visits combined from 6 months after dual therapy initiation to treatment change or the end of follow-up, with secondary analyses of individual components. Rate ratios were calculated using negative binomial regression with adjustment for propensity scores.

### *Results*

1,704 patients were included in the propensity score matched cohort. There was weak evidence of increased risk of secondary healthcare admissions or visits [adjusted rate ratio 1.12, 95% confidence interval (CI) 0.97-1.29]. Evidence of increased risk was strongest for inpatient admissions [adjusted rate ratio 1.38 (95% CI 0.95-2.00)], and specifically for macrovascular admissions which accounted for 77.2% of inpatient admissions [adjusted rate ratio 1.77 (95% CI 1.15-2.71)].

### *Conclusions*

Choice of second line anti-hyperglycaemic agent appears to increase secondary healthcare admissions, especially for cardiovascular disease. This adds to existing evidence that health economic outcomes are important considerations for T2DM treatment decisions and demonstrates the benefits of using linked hospital datasets to measure health economic outcomes related to inpatient and outpatient hospitalisations.

## *Paper 3*

Pioglitazone use and risk of bladder cancer in patients with type 2 diabetes: retrospective cohort study using datasets from four European countries (Korhonen et al, 2016[8])

### *Introduction*

Pioglitazone is an oral anti-hyperglycaemic drug indicated as a monotherapy or dual oral therapy in patients with T2DM inadequately controlled or with contraindications to metformin or sulphonylureas monotherapy[59]. In 2005, an

increased risk of bladder cancer was observed in patients treated with pioglitazone compared to placebo in the prospective pioglitazone clinical trial in macrovascular events (PROactive)[60]. This association was supported by subsequent epidemiology studies[61] resulting in the addition of product warnings[59] and requests for rigorous large-scale observational studies with a focus on minimising channelling bias.

### *Objective*

To estimate absolute and relative risks of bladder cancer with use of pioglitazone compared to alternative treatment regimens in patients with T2DM.

### *Methods*

We conducted a retrospective cohort study in patients with T2DM recorded in linked national healthcare databases in Finland and Sweden, PHARMO databases in the Netherlands, and CPRD data in the UK. The UK cohorts were divided into linked and unlinked cohorts according to linkage eligibility and general practice and hospital cohorts were included in the Netherlands. The linked UK cohort included CPRD GOLD, HES APC, cancer registration and ONS mortality data. Patients in each cohort who initiated pioglitazone were matched with patients with T2DM who had never received pioglitazone by treatment stage, history of diabetes, diabetes complications, cardiovascular disease and year of cohort entry using a combination of propensity score and individual variable matching. Country specific cohorts were pooled to create a pan-European analysis cohort. Crude incidence rates were calculated in pioglitazone exposed and comparator groups. Cox proportional hazards models were used to estimate hazard ratios with adjustment for baseline and time dependent covariates.

### *Results*

In the primary 1:1 matched analysis, 56,337 pioglitazone exposed patients and never exposed controls were followed up for a mean 2.9 and 2.8 years respectively during which 130 and 153 bladder cancers were recorded. The crude incidence of bladder cancer was 7.97 per 10,000 person years in the pioglitazone exposed group and 9.62 per 10,000 patient years in the never exposed group. The adjusted hazard ratio (HR) for patients ever exposed versus never exposed to pioglitazone was 0.99 (95% CI 0.75-1.30). Duration of use and cumulative dose were not associated with risk of bladder cancer. In stratified analyses,



heterogeneity was observed between country and datatype cohorts with HRs ranging from 0.56 (95% CI 0.31 to 1.00) in Finland to 4.27 (1.26- 14.46) in Sweden.

### *Conclusions*

The findings from this pooled analysis of linked databases from four European countries are consistent with the absence of a causal association between exposure to pioglitazone and bladder cancer in patients with T2DM. Since publication, this study has been included in two meta-analyses of observational research studies; both concluded that there was evidence of a possible association with adjusted risk estimates of 1.16 [95% CI 1.04-1.28][62] and 1.13 [95% CI 1.03- 1.25][63].

### *Papers 4 and 5*

Pioglitazone and risk of mortality in patients with type 2 diabetes: results from a European multidatabase cohort study (Strongman et al, 2017[9])

Pioglitazone and cause-specific risk of mortality in patients with type 2 diabetes: extended analysis from a European multidatabase cohort study (Strongman et al, 2019[10])

### *Introduction*

The long-term goal of diabetes treatment is to reduce microvascular and macrovascular events and related morbidity and mortality but RCTs of anti-diabetes drugs are not typically powered to detect differences in mortality between arms. A meta-analysis largely driven by two RCTs in type 2 diabetic patients at high risk of cardiovascular disease (PROactive) and prediabetic patients with a history of ischemic stroke or transient ischemic attack (Insulin Resistance Intervention after Stroke trial) did not find evidence of an association between pioglitazone use and risk of death (RR 0.93, 95% CI 0.80-1.09) [64]. In contrast observational studies indicate substantial reductions in all-cause mortality with pioglitazone use compared to insulin (HR 0.33, 95% CI 0.31-0.36)[65] and non-use of pioglitazone (HR 0.77, 95% CI 0.71-0.84)[66].

### *Objective*

Paper 4: To estimate absolute and relative risks of all-cause mortality in patients whose T2DM therapy is changed to include pioglitazone versus an alternative antidiabetic regimen at the same stage of disease progression.

Paper 5: Exploratory analysis of cardiovascular and non-cardiovascular mortality in the same cohort

### *Methods*

Crude mortality and hazard ratios were generated using the matched cohort and exposure definitions and statistical methodology from the primary bladder cancer analysis (Paper 3). In Finland, Sweden, The Netherlands hospital and UK linked datasets, national death registration data were used to measure all-cause mortality; cause-specific mortality was also measured except in the Netherlands. General practice records were used to measure all-cause mortality in the UK and Netherlands GP databases.

### *Results*

Paper 4: 3,370 and 7,143 deaths occurred in the pioglitazone exposed and unexposed groups respectively over 2.9 and 2.8 years of follow-up. Crude mortality rates per 10,000 patient years were 206 (95% CI 199- 213) for patients ever exposed to pioglitazone and 448 (95% CI 438- 458) for patients never exposed to pioglitazone. In unadjusted and adjusted analyses, a 54% (95% CI 52- 55) and 33% (95% CI 30- 36) reduction in risk of mortality was observed with pioglitazone use. A reduction in risk was observed in all countries and datatypes with considerable variation in effect size in the adjusted analysis (11% in the Netherlands general practice to 46% in Finland).

Paper 5: Substantial reductions in both cardiovascular and non-cardiovascular mortality were observed when comparing ever versus never exposure to pioglitazone: cardiovascular HR 0.58 (95% CI 0.52-0.63); non-cardiovascular HR 0.63 (95% CI 0.58-0.68).

### *Conclusion*

Patients who are prescribed pioglitazone have lower mortality risks than patients prescribed alternative anti-diabetic treatments at a similar stage of disease progression. Using linked national death registration databases, these reductions

in risk were observed for both cardiovascular and non-cardiovascular mortality. Caution should be applied in interpreting this as a causal association as substantial reductions in risk were not observed in RCTs and this study was primarily designed to investigate bladder cancer risk.

## Discussion

### Key findings

CPRD link primary care data to secondary care data collected by the NHS and area-based measures of deprivation through a well-governed, robust and resource-saving centralised framework. The processes are described in Paper 1 which can be used to inform the design of applied studies, understanding of potential limitations, and future linkage projects internationally. These research implications and the strengths and limitations of the processes and paper are discussed in relation to existing literature below.

Papers 2 to 5 provide applied examples in the field of diabetes pharmacoepidemiology demonstrating the value of these data to compare safety, mortality and healthcare resource utilisation outcomes for different anti-diabetic drug exposures. These studies involve multiple decisions and assumptions to identify study populations, exposures, outcomes and covariates; and to conduct robust statistical analyses that compensate for the limitations of observational data. The strengths and limitations of these decisions and assumptions are described in the context of published validation, concordance and methodological papers in my discussion related to objectives 2 to 5 below.

### Objective 1

*Understand the methods used by NHS Digital to link CPRD primary care data to other health data sources, the processing steps implemented by CPRD, and the implications of these methods for study design and reporting. Update these processes and increase transparency and awareness of these issues by publishing a peer-reviewed manuscript.*

CPRD GOLD data are linked to secondary healthcare datasets through a trusted third party, NHS Digital. These data widen the range of exposures, outcomes and covariates that can be identified for applied research, and improve validity of measurements, especially for diseases that are treated in multiple settings[36–38,67,68]. Data do not need to be linked separately for individual studies, concentrating resources to ensure robust processing and trustworthy use of data. Although they are not involved in the linkage process, applied researchers need to

understand these processes and their implications for study design and potential research limitations.

CPRD and NHS Digital use a stepwise deterministic linkage method centred on a complete identifier, the NHS number, to link CPRD primary care data to secondary linked datasets. The processes followed maintain separation of anonymised clinical and personal identifiable data respecting patients' rights to privacy and UK law. Applied researchers should use metadata provided with linked cohorts to restrict denominator populations to patients included in the linkage process and analyses to time periods covered by all linked datasets. Without this information, bias may be introduced through differential misclassification of study variables in patients who were not eligible for linkage or during time periods that are not covered by the linked data[69]. Transparency could be further improved through a comparison of deterministic and probabilistic methodologies by NHS Digital and CPRD. Continued efforts need to be made to ensure that researchers using these data describe and justify how they use meta-data to identify study populations.

## Objective 2

*Demonstrate general benefits, limitations and best practice in using primary care and linked health data in the field of diabetes pharmacoepidemiology.*

### *Identifying patients with T2DM*

Although oral antihyperglycaemic drugs are indicated to treat T2DM, they are sometimes also used for T1DM[70] and other indications such as polycystic ovary syndrome[71]. To exclude patients with alternative indications, study populations for T2DM pharmacoepidemiology studies are therefore commonly restricted to patients with a record for T2DM. This approach was followed for the pioglitazone studies (Papers 3 to 5); a record of metformin prescribing was sufficient to include patients in Paper 2. Both studies were restricted to patients over 40 at diabetes diagnosis. As unlicensed use of the later stage T2DM regimens included in these studies is unlikely, these steps may have unnecessarily excluded patients reducing generalisability of the studies.

Study populations that are not defined by treatments that are largely exclusively used in type 2 diabetes are at greater risk of misclassification of type 2 diabetes due to lack of specificity of coding. These include studies identifying diabetes as

an exposure or outcome or including patients treated with diet and exercise alone or earlier stage oral treatment regimens; more rigorous definitions of T2DM should be used for these studies[30–33].

#### Identification of patients with other medical conditions

CPRD GOLD, HES APC, NCRAS and ONS mortality data were used to identify the bladder cancer outcome in the linked UK dataset for Paper 3. Reported concordance of recording between sources for urinary tract cancer suggests that this maximises identification of true cases of bladder cancer (sensitivity) but may have introduced a small proportion of false positive outcomes[72]. Assuming little difference between patients in the linked and unlinked datasets, this explains the higher observed incidence of bladder cancer in the linked cohort. This increase was observed in both the exposed and unexposed groups but is particularly high in the nearest matched unexposed group. If this is the result of differential misclassification, it may partly explain observed heterogeneity in the hazard ratios between the two cohorts (see objective 4). Stage and grade of bladder cancer was recorded in too few patients to be of use for this study.

Linked data sources were also used to identify baseline covariates in both studies increasing ascertainment of conditions that are commonly treated in secondary care such as cardiovascular disease.

#### Developing treatment algorithms

Algorithms were used to identify exposed and comparison groups for Paper 2 (date of addition of SU or alternative antidiabetic agent to first line metformin therapy); duration of continuous prescribing for both studies; and cumulative dose, time since last dose and patterns of treatment change for Papers 3 to 5. These are described in the Web appendix (Paper 2) and protocol appendix 2 (Papers 3 to 5)[73].

These algorithms bring fractured prescription records together using assumptions to impute missing or implausible quantities and daily doses and to account for gaps and overlaps in prescribing. These assumptions have not been standardised or validated. More recently, these techniques have been described mathematically[40].

National patient level hospital prescribing data are not available in the UK. Linked HES APC and HES outpatient data could be used to identify long-term inpatient admissions or outpatient visits that may be associated with gaps in prescribing or treatment switches.

In Paper 2, I used measures of continuous prescribing to calculate the MPR, a ratio of the number of days covered by prescriptions over the number of days observed in the dual therapy period. A low proportion of patients had an MPR less than 80% in either cohort (4.4% SU, 2.9% OHA). Lack of compliance may be more strongly related to drug dispensing than prescribing and whether the patient took the drug once dispensed. This cannot be measured in CPRD primary care data. Alternative, less commonly used, measures of compliance are the Continuous measure of Medication Gap (CMG) which represents the percentage of time that the patient does not have the medication available[41] and the maximum medication gap[74]. A Swedish study identified higher level of adherence using the Continuous measure of Medication Acquisition (CMA, adherent if CMA  $\geq$  80 %) (adherent if gaps <45 days)[74]. The CMA is similar to the MPR. The direction of relative differences between comparator drug groups were the same for each method. I prefer the MPR as this takes prescribing behaviour over a longer time period into account and seems less vulnerable to individual anomalies in prescribing behaviour.

#### [Measuring additional covariates and accounting for missing data](#)

I used recommendations in published research to guide assumptions about the missingness mechanism for Paper 2[44]. Smoking status was assumed to be more likely to be missing in non-smokers than smokers in UK primary care data i.e. missing not at random. Smoking status categories were therefore changed to current, past and no evidence of smoking. Missingness was assumed to be at random for other variables; values were estimated through multiple imputation. This assumption is unlikely to be true as the covariates themselves are likely to influence healthcare reporting (e.g. underweight patients more likely to have BMI recorded). This is unlikely to have influenced the study findings due to low levels of missing data (minimum 0.3% for smoking status, maximum 2.4% for HbA1c)

Missing values for smoking, BMI and HbA1c were treated as a separate category in Papers 3 to 5. This commonly used methodology may have biased the effect

measures as it has been shown to lead to substantial inaccuracies under alternatives missing data assumptions[75], and high levels of missingness were observed. The 18.7% change in the hazard ratio in datasets in which these variables can be measured from 1.02 (95% CI 0.70-1.49) in the original model to 0.83 (95% CI 0.54-1.28) when these variables were included is therefore difficult to interpret.

Complete case analysis may have been more appropriate for both studies, whereby patients with incomplete records are excluded from analyses. Assuming that the reasons for missingness are not associated with the outcome conditional on measured covariates [76], this method is unlikely to cause bias.

#### Minimising confounding and channelling bias

Propensity scores were used to address the risk of channelling bias in all applied studies.

These were estimated using logistic regression models including baseline covariates that were assumed to influence both the exposure and outcome. Caliper matching was used to balance the characteristics of exposed and unexposed groups based on covariates included in the score. This has the additional advantage of restricting patients included in the comparison to those with characteristics that led to differential prescribing[51]. Good balance was achieved in variables included in the propensity score as demonstrated in descriptive tables comparing exposure groups in all papers. However, there is little evidence that propensity scores improve balance for unmeasured covariates[51] and traditional techniques may have been equally valid, especially for Paper 2 where the outcome was not rare leading to minimal risk of difficulties with maximum likelihood estimation in regression models[77]. The choice of methodology for both studies was strongly influenced by the preference for propensity scores by drug regulators including the European Medicines Agency (EMA). As the researchers in our pan-European pioglitazone panel are less convinced, mixed methodologies were used for our analyses. Covariates considered to be associated more closely with the outcome than exposure were not included in propensity scores but were adjusted for in final statistical analyses, and exact matching was required for the three main propensity score variables. Two cohorts were defined: a 1:1 nearest match cohort and an up to 1:10 matched



cohort. As matching ratios varied between patients and countries in the multiple matched cohort, balancing weights were used in Cox proportional hazards models and when calculating standardised differences[78]. Inclusion of all variables in the propensity score would have added clarity to the design, and provided that the models were correctly specified, would have minimised channelling bias at least as effectively as the mixed method.

#### [Avoiding the use of future information and time-related biases](#)

All papers selected exposed and comparator groups and started follow-up at the same stage of T2DM treatment. Exclusion criteria relating to follow-up time prior to exposure were the same in exposed and comparator groups. Time-related biases such as immortal time bias that result from differential inclusion or exclusion of follow-up time in the exposed and comparison groups were therefore largely avoided.

Follow up metformin and SU/alternative oral anti-diabetes agent prescriptions were required to confirm dual therapy prescribing in Paper 2. This would have been an example of use of future information had follow-up started at the start of dual therapy. Patient time from index date to the end of follow-up would have been excluded in patients with insufficient follow-up time for these confirmatory prescriptions to have been captured. This would cause bias if the association between having sufficient follow-up and the outcome differed in exposed and comparator patients. This type of bias was avoided by starting follow-up 6 months after the index date. This method has the additional advantages of avoiding protopathic bias where appearance of early symptoms of the outcome under investigation prompts changes in treatment and accounting for hypothesized lags in treatment effects.

#### [Summary](#)

Primary care data can be used to identify patients with diabetes, treatment exposures and medical observations measured in primary care. These data are used to measure associations between related outcomes and exposures and to estimate causal effects through adjustment for known and measured confounders. Use of linked primary and other health datasets improves ascertainment of conditions that are also treated in secondary care such as bladder cancer.

Researchers make multiple assumptions when bringing event records together to measure study variables and design statistical analyses plans. Validation and methodological studies are important guides but decisions should be made within the context of the individual study and limitations recognised.

### Objective 3

*Demonstrate the benefits and limitations of using linked HES APC and outpatient data to measure healthcare resource utilisation in the field of diabetes pharmacoepidemiology*

Collection of HES APC data was established in 1989/1990, with mandatory inclusion of NHS number since 1997/1998. Each hospitalisation is divided into episodes describing care under a single consultant with associated ICD-10 codes identifying the primary diagnosis accounting for most of the length of stay and secondary conditions or comorbidities.[28] In Paper 2, primary ICD-10 codes for the first episode of a hospitalisation were used to identify diabetes-associated admissions for macrovascular, microvascular, hypoglycaemia, diabetes-coded admissions and falls. Hospital admissions with OPCS codes for amputation were also included as macrovascular admissions.

HES outpatient data are available separately with collection beginning in April 2003. Data were considered exploratory until 2007 due to incomplete coverage for some specialists and mandatory collection of nurse and midwife appointments starting in April 2005. Recording of diagnosis and procedure codes is not mandatory; they were recorded in 1.6% and 2.6% of appointments in 2004/2005 and 4.9% and 30% of appointments in 2016/2017.[79,80] I therefore set the start of our study period to 2003, acknowledging potential for missing data in the early years, and used specialist codes to identify diabetes-associated outpatient visits.

The adjusted rate ratio for all diabetes-related hospital visits combined was 1.12 (95% CI 0.97-1.29) providing weak evidence of increased diabetes-related hospital use by patients prescribed SUs as a second line add-on to metformin compared to other oral anti-diabetic drugs. The rate ratio was higher for inpatient admissions than outpatient visits and substantially greater for macrovascular admissions and outpatient visits to cardiology. Outpatient visits to the range of specialties included in our study are less likely to be diabetes-specific than inpatient

admissions restricted to diabetes-related conditions. This assumption also applies to cardiology visits and admissions compared to other specialties such as general medicine. Our observation of higher rate ratios for inpatient visits and cardiology compared to our combined primary outcome is therefore consistent with the theory that non-differential misclassification moves the estimate towards the null[81]. This illustrates a limitation of using HES data, especially outpatient data, to identify diabetes-associated hospital visits.

Secondary outcomes available in the HES APC data were route of admission (A&E, non-A&E) and length of stay. I could not count all A&E attendances as these data were not available at the time of the study.

I estimated inpatient and outpatient costs through linkage to Payment-by-results tariffs through the Healthcare Resource Groups (HRG4) grouper and tariffs for consultant visits[82].

Overall this study demonstrates the strengths of linkage to HES datasets to measure a range of healthcare resource utilisation outcomes that are inconsistently recorded in primary care data. Strong assumptions need to be made when attributing visits to a specific cause such as diabetes, especially for the outpatient data. Potential coding errors in the source data and linkage errors both within the HES datasets and in linking to CPRD should also be acknowledged.

## Objective 4

*Demonstrate the challenges, benefits and limitations of pooling linked data from different countries to assess a rare cancer outcome in the field of diabetes pharmacoepidemiology*

Combining data from multiple countries in a single pre-planned study is increasingly popular in pharmacoepidemiology, mainly motivated by increased sample size and more precise estimates where exposures or outcomes are rare[83]. Researchers either combine aggregate results from individual country analyses using standard meta-analytical techniques or pool individual data from each country. The latter approach was used to combine data from six data sources from four European countries for Papers 3 to 5. A single protocol and statistical analysis plan were developed for the study including standardised

exposure, outcome and covariate definitions (where possible) and a single analysis approach. Analysis datasets were created for individual countries and pooled in a single dataset. Primary analyses were based on the pooled dataset which included a country/source variable. To investigate heterogeneity between countries and sources, secondary analyses were stratified by country and fixed and random effects meta-analyses completed.

The adjusted hazard ratio for bladder cancer in patients ever exposed to pioglitazone in the nearest match cohort was 0.99 (95% CI 0.75-1.30). Confidence intervals were wide for individual country analyses reflecting low numbers of bladder cancer outcomes, especially in Sweden (Paper 3, Figure 3). Fixed and random effects meta-analyses resulted in very similar results, especially in the multiple matched cohort, and identified statistically significant heterogeneity between countries (Paper 3, Figure 4).

Potential systematic differences between countries and data sources include differences between patients with T2DM between countries that are associated with bladder cancer outcomes and were not accounted for by our matching or statistical analysis plans. Differential prescribing of pioglitazone (e.g. higher doses in different countries) may also lead to different hazard ratios. Use of different data sources to identify exposures, outcomes and covariates will also impact the models. In sensitivity analyses, the HR for datasets including cancer registration data was below one whereas the HRs for datasets without this data was above one. This is mirrored in differences between the UK linked and unlinked datasets and may be explained by improved classification of the bladder outcome with the use of cancer registration data. All datasets are affected by left truncation of data which makes it difficult to differentiate between incident and prevalent prescribing, and to identify variables such as duration of diabetes and previous treatments prescribed. This misclassification was strongest in Sweden where pioglitazone has been available since 2001 but prescribing data collection began in 2005. Finally, although analytical datasets were created using a common statistical analysis plan, there may have been differences in interpretation between analysts in each country. One known example of this which is reported in the supplementary appendix is that for those who were never exposed to pioglitazone, the first initiations of alternative new antidiabetic drugs were potential cohort entry dates in UK and Netherlands datasets, whereas all changes

in the antidiabetic drug treatment were potential cohort entry dates in Finland and Sweden datasets. This affected the size and composition of the matching pool in different countries.

In summary, use of pooled linked data sources from multiple data sources allowed us to obtain precise estimates for a rare outcome in diabetes pharmacoepidemiology. Differences in effects between countries were observed. These are likely to be influenced by systematic differences in diabetes patients and treatments between countries and differences in recording between datasets. A further limitation is loss of information from individual databases due to the need to standardise. For example, the main analysis did not include HbA1c and smoking status as these data were not available in all countries. This may lead to residual confounding.

## Objective 5

*Demonstrate the challenges, benefits and limitations of using linked data to assess all-cause and cause-specific mortality outcomes in the field of diabetes pharmacoepidemiology*

All-cause mortality data were available for Paper 4 in all countries and datasets from either national death registries or general practice records. Our cause-specific mortality analysis (Paper 5) was restricted to datasets that included linked national death registration data that included this information. National cause of death registries follow internationally agreed rules to assign World Health Organisation ICD-10 codes as underlying and contributing causes of death[84]. Nevertheless, these depend on the quality of underlying medical certification by clinicians [85] and discrepancies have been observed between clinical and autopsy diagnoses; for example, pulmonary embolism, ischaemic heart disease/myocardial infarction and pneumonia are often confused with each other[86]. Differences have been observed between countries including assignment of diabetes as a contributing or underlying cause[85].

All-cause and cause-specific mortality were secondary and exploratory outcomes for a study whose primary aim was to assess the association between pioglitazone prescribing and bladder cancer. Our comparison group definitions and statistical analysis techniques focussed on channelling and other biases that might mask or

distort the association between pioglitazone prescribing and bladder cancer. If the study had been specifically designed to study all-cause mortality, more attention would have been paid to reverse causality due to diagnoses of life limiting diseases such as cancer prior to pioglitazone prescribing and to confounding by socio-economic status which may have a stronger association with mortality than bladder cancer[87]. Sensitivity analyses for the cause-specific mortality analysis (Paper 5) identified substantial heterogeneity between countries and between potential effect modifiers whose prevalence varies between countries such as history of diabetes complications, chronic kidney disease and history of thiazolidinediones use at cohort entry. The influence of differences between populations and recording of confounders in different countries are therefore difficult to disentangle.

These complications led us to recommend further studies that are specifically designed to study mortality outcomes. Ideally, studies would be designed separately for individual countries to allow inclusion of important covariates where available. In linked CPRD studies, these would include HbA1c, smoking status and BMI. These were included in a sensitivity analysis restricted to UK and Swedish study in Paper 5, reducing the observed protective effects from HR 0.8 (95% CI 0.69-0.93) to HR 0.85 (0.72-1.01) for cardiovascular mortality with similar findings for non-cardiovascular mortality. It would also be important to include area-based socio-economic status measures which are available in the UK.

### Implications for future research

Paper 1 describes the process used to link primary care and other health data sources. Further methodological research is needed to increase transparency of the five-step algorithm used by NHS Digital to link health data, understand potential biases introduced through the linkage process and validate CPRD's approach to restrict data to records linked at steps 1 to 5 of the linkage algorithm. Within linked concordance studies, possible approaches to improve understanding of these potential biases and validate CPRD's approach include comparing measures of agreement between patients matched at different steps of the algorithm. Depending on information governance considerations, linkage may also introduce the opportunity to join records for individuals that have registered in more than one CPRD practice, extending individual follow-up time. When patients join a new GP practice, data from their previous record may now

be transferred electronically to the new practice[88], a form of linkage in itself. It may also therefore be possible to extend individual follow-up to before registration in the current practice. To do this, we would either need metadata indicating when practices implemented GP2GP transfer and in which patients, or to identify the start of relevant data collection for an individual study. For example, studies investigating the effects of later stage diabetes drugs such as pioglitazone could start follow-up from the first recorded laboratory test or prescription for each patient. This assumes that retrospective records of these data would only be added to a patient record prospectively or by GP2GP transfer and that patients with diabetes would have at least one such record around the time of diagnosis allowing researchers to measure variables describing stage of disease progression and treatment. Changes in electronic data transfer and General Practice management also have implications for the CPRD derived up to standard date which is used to determine the latest date from which complete data recording is available for individual practices. CPRD have not yet published an up to standard date for the CPRD Aurum data making this an ideal time for the wider research community to contribute. The starting point could be qualitative research with GPs, practice managers and software providers to assess drivers of gaps in data collection within practices and events that may have influenced current UTS dates in CPRD GOLD practices.

Before undertaking a new study, researchers should systematically review existing published and validated definitions of their main exposures and outcomes and consider novel validation studies where evidence is lacking or out of date.

Examples that I have identified include:

- Using machine learning methodologies to recognise and identify episodes of continuous prescribing and changes of diabetes drugs and regimens.
- Exploring potential biases related to differential and non-differential misclassification of different diabetes drugs and regimens.
- Estimating the impact of using prescribing data compared to dispensing data using data from countries where both of these sources are available, including Scotland[89].
- Validating recording of underlying and direct causes of death in death registration data in different countries and understanding the potential impact of this on research.

Methodological research and reviews are also needed to increase confidence in choosing between multivariable regression models and propensity score methods; and to demonstrate the benefits and limitations, and validate the assumptions, of commonly used methods to deal with missing data in different scenarios.

The observed differences between countries in the pooled pioglitazone analyses, and difficulties in understanding the reasons for this, have implications for the design and conduct of future pooled pharmaco-epidemiology studies. One key recommendation that arises from this is that when planning studies, investigators formally describe how selected databases differ in terms of the nature and size of the source population, treatment guidelines, and likely prevalence and recording of the exposure, outcome and important confounders and effect modifiers. Investigators should use this information to assess whether the databases are suitable for a pooled analysis and to plan sensitivity analyses that explore the impact of these differences within and between countries and datasets. To avoid differences in interpretation of the study protocol, common programmes should be developed and applied to individual databases where possible. Before starting statistical analyses, participant flow charts and baseline characteristics should be compared and unexpected differences between countries investigated. This would increase confidence that the protocol has been interpreted in the same way in each dataset and further understanding of differences between countries.

### Implications for policy

This thesis has described an example framework for providing linked primary care and health data for research, and demonstrated the benefits of using primary care and linked health data in the field of diabetes pharmacoepidemiology. Until recently, these data have been underused by HTA associations such as NICE[90]. Recent strategies of the EMA[91], MHRA[92] and NICE[93] to increase use of these data, whilst taking their limitations into account, are warranted. Current CPRD data linkages are ideal for the study of pharmacoepidemiology and the UK government has proposed new initiatives to extend collection of these data. These initiatives need to fully understand the complex information governance environment and ensure that plans for data collection continue to support a wide range of epidemiological studies[94].



Data availability should be extended beyond routinely collected NHS data to support the full range of research that is required to inform prevention and treatment of diabetes, a disease that is influenced as much by our environment as factors recorded within the health care system[95]. These data are available in prospective cohort studies and non-health administrative datasets. Linkage to these data has been hindered by the lack of a clear legal route to enable linkage, inconsistent understanding of legislation and guidelines by different government departments and agencies, and the need to demonstrate that research is in the public interest and supported by patients[96]. Organisations such as Health Data Research UK[97] have a key role in overcoming these hurdles.

## Conclusions

CPRD primary care and linked data are a useful resource for pharmacoepidemiology including in the field of diabetes. Centralised linkage of records through a standardised linkage algorithm following robust information governance procedures supports trustworthy and resource efficient use of these data. Multiple decisions and assumptions are required at all stages of applied study design. These include selecting the optimal combination of data sources, selecting the study population, identifying study variables, and choosing statistical analyses methodologies. Published methodology studies support decision making and increase confidence in research. Nevertheless, one size does not fit all; epidemiologists need to carefully consider each decision in the context of the data available and their research question.

## References

- 1 Langan SM, Schmidt SA, Wing K, *et al.* The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ* 2018;**363**:k3532.
- 2 Farmer R, Mathur R, Bhaskaran K, *et al.* Promises and pitfalls of electronic health record analysis. *Diabetologia* 2018;**61**:1241–8.
- 3 Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes Care* 2012;**35**:2665–73.
- 4 World Health Organisation (WHO). Global report on diabetes. *WHO* Published Online First: 2016.<https://www.who.int/diabetes/global-report/en/> (accessed 3 May 2019).
- 5 Vezyridis P, Timmons S. Evolution of primary care databases in UK: a scientometric analysis of research output. *BMJ Open* 2016;**6**:e012785.
- 6 Padmanabhan S, Carty L, Cameron E, *et al.* Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *Eur J Epidemiol* 2019;**34**:91–9.
- 7 Strongman H, D’Oca K, Langerman H, *et al.* Comparison of diabetes-associated secondary healthcare utilization between alternative oral antihyperglycaemic dual therapy combinations with metformin in patients with type 2 diabetes: An observational cohort study. *Diabetes, Obes Metab* 2015;**17**.
- 8 Korhonen P, Heintjes EM, Williams R, *et al.* Pioglitazone use and risk of bladder cancer in patients with type 2 diabetes: retrospective cohort study using datasets from four European countries. *BMJ* 2016;**354**:i3903.
- 9 Strongman H, Korhonen P, Williams R, *et al.* 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**:e000364.
- 10 Strongman H, Christopher S, Majak M, *et al.* Pioglitazone and cause-specific risk of mortality in patients with type 2 diabetes: extended analysis

from a European multidatabase cohort study. *BMJ Open Diabetes Res Care* 2018;**6**:e000481.

- 11 National Institute for Health and Care Excellence. Type 2 diabetes: newer agents for blood glucose control in type 2 diabetes. 2009.<http://www.nice.org.uk/guidance/cg87/resources/cg87-type-2-diabetes-newer-agents-a-partial-update-of-cg66-short-guideline2> (accessed 27 Aug 2014).
- 12 United Nations General Assembly. Resolution 61/225. World Diabetes Day. 2007.
- 13 National Institute for Health and Care Excellence (NICE). Overview | Type 2 diabetes in adults: management | Guidance. <https://www.nice.org.uk/guidance/ng28> (accessed 10 May 2019).
- 14 National Institute for Health and Care Excellence (NICE). Overview | Type 1 diabetes in adults: diagnosis and management | Guidance. Published Online First: 2016.<https://www.nice.org.uk/guidance/ng17> (accessed 14 Jul 2019).
- 15 NHS Digital. Patients Registered at a GP Practice. <https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice> (accessed 12 Jan 2018).
- 16 Office for National Statistics. Population estimates. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates> (accessed 25 Apr 2019).
- 17 National Audit Office Report (HC 192 2012-13): Healthcare across the UK: A comparison of the NHS in England, Scotland, Wales and Northern Ireland. 2012. <https://www.nao.org.uk/wp-content/uploads/2012/06/1213192.pdf> (accessed 25 Apr 2019).
- 18 Benson T. Why general practitioners use computers and hospital doctors do not--Part 1: incentives. *BMJ* 2002;**325**:1086–9.<http://www.ncbi.nlm.nih.gov/pubmed/12424171> (accessed 25 Apr 2019).

- 19 Geoffrey Rivett. Chapter 6 Labour's Decade 1998 - 2007. In: *National Health Service History*.  
[http://www.nhshistory.net/chapter\\_6.html#\\_ednref6](http://www.nhshistory.net/chapter_6.html#_ednref6) (accessed 25 Apr 2019).
- 20 GP Systems of Choice - NHS Digital. <https://digital.nhs.uk/services/gp-systems-of-choice> (accessed 14 Jul 2019).
- 21 Benson T. The history of the Read Codes: the inaugural James Read Memorial Lecture 2011. *Inform Prim Care* 2011;**19**:173–82.  
<http://www.ncbi.nlm.nih.gov/pubmed/22688227> (accessed 25 Apr 2019).
- 22 SNOMED CT implementation in primary care - NHS Digital.  
<https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct/snomed-ct-implementation-in-primary-care> (accessed 14 Jul 2019).
- 23 Healthcare Data and Research2 | THIN Data. <https://www.the-health-improvement-network.co.uk/> (accessed 14 Jul 2019).
- 24 Home - QResearch. <https://www.qresearch.org/> (accessed 14 Jul 2019).
- 25 Chaudhry Z, Mannan F, Gibson-White A, *et al*. Outputs and Growth of Primary Care Databases in the United Kingdom: Bibliometric Analysis. *J Innov Heal Informatics* 2017;**24**:284.
- 26 Wolf A, Dedman D, Campbell J, *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* Published Online First: 11 March 2019.
- 27 Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;**44**:827–36.
- 28 Herbert A, Wijlaars L, Zylbersztejn A, *et al*. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017;**46**:1093-1093i.
- 29 Henson KE, Elliss-Brookes L, Coupland VH, *et al*. Data Resource Profile: National Cancer Registration Dataset in England. *Int J Epidemiol* Published Online First: 23 April 2019.

- 30 Eastwood S V, Mathur R, Atkinson M, *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLoS One* 2016;**11**:e0162388.
- 31 Mathur R, Bhaskaran K, Edwards E, *et al.* Population trends in the 10-year incidence and prevalence of diabetic retinopathy in the UK: a cohort study in the Clinical Practice Research Datalink 2004–2014. *BMJ Open* 2017;**7**:e014444.
- 32 Holden SH, Barnett AH, Peters JR, *et al.* The incidence of type 2 diabetes in the United Kingdom from 1991 to 2010. *Diabetes Obes Metab* 2013;**15**:844–52.
- 33 Tate R, Dungey S, Glew S, *et al.* Quality of recording of diabetes in the UK: how does the GP’s method of coding clinical data affect incidence estimates? Cross-sectional study using the CPRD database. *BMJ Open* 2016;**7**:e012905.
- 34 National Institute For Health and Care Excellence (NICE). Standards and Indicators. <https://www.nice.org.uk/standards-and-indicators?tab=qof> (accessed 10 May 2019).
- 35 Herrett E, Thomas SL, Schoonen WM, *et al.* Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 2010;**69**:4–14.
- 36 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.
- 37 Crooks CJ, Card TR, West J. Defining upper gastrointestinal bleeding from linked primary and secondary care data and the effect on occurrence and 28 day mortality. *BMC Health Serv Res* 2012;**12**:392.
- 38 Arhi CS, Bottle A, Burns EM, *et al.* Comparison of cancer diagnosis recording between the Clinical Practice Research Datalink, Cancer Registry and Hospital Episodes Statistics. *Cancer Epidemiol* 2018;**57**:148–57.
- 39 Nielsen LH, Løkkegaard E, Andreassen AH, *et al.* Using prescription registries

- to define continuous drug use: how to fill gaps between prescriptions. *Pharmacoepidemiol Drug Saf* 2008;**17**:384–8.
- 40 Khotimah PH, Sugiyama Y, Yoshikawa M, *et al.* Medication Episode Construction Framework for Retrospective Database Analyses of Patients With Chronic Diseases. *IEEE J Biomed Heal Informatics* 2018;**22**:1949–59.
- 41 Peterson AM, Nau DP, Cramer JA, *et al.* A Checklist for Medication Compliance and Persistence Studies Using Retrospective Databases. *Value Heal* 2007;**10**:3–12.
- 42 Booth HP, Prevost AT, Gulliford MC. Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for England, 2007 to 2011. *Pharmacoepidemiol Drug Saf* 2013;**22**:1357–61.
- 43 Bhaskaran K, Forbes HJ, Douglas I, *et al.* Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). *BMJ Open* 2013;**3**:e003389.
- 44 Marston L, Carpenter JR, Walters KR, *et al.* Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf* 2010;**19**:618–26.
- 45 Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? *Int J Epidemiol* 2014;**43**:1336–9.
- 46 Sterne JAC, White IR, Carlin JB, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;**338**:b2393.
- 47 Saine ME, Carbonari DM, Newcomb CW, *et al.* Concordance of hospitalizations between Clinical Practice Research Datalink and linked Hospital Episode Statistics among patients treated with oral antidiabetic therapies. *Pharmacoepidemiol Drug Saf* 2019;**28**:1328-1335.
- 48 Harshfield A, Abel GA, Barclay S, *et al.* Do GPs accurately record date of death? A UK observational analysis. *BMJ Support Palliat Care* 2018; Published Online First: 27 June 2018. doi: 10.1136/bmjspcare-2018-001514
- 49 Kabisch M, Ruckes C, Seibert-Grafe M, *et al.* Randomized Controlled Trials:

Part 17 of a Series on Evaluation of Scientific Publications. *Dtsch Arztebl Int* 2011;**108**:663.

- 50 Brookhart MA, Stürmer T, Glynn RJ, *et al.* Confounding control in healthcare database research: challenges and potential approaches. *Med Care* 2010;**48**:S114-20.
- 51 Glynn RJ, Schneeweiss S, Sturmer T. Indications for Propensity Scores and Review of their Use in Pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006;**98**:253–9. doi:10.1111/j.1742-7843.2006.pto\_293.x
- 52 Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
- 53 Pacurariu A, Plueschke K, McGettigan P, *et al.* Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open* 2018;**8**:e023090.
- 54 Morgan CL, Mukherjee J, Jenkins-Jones S, *et al.* Combination therapy with metformin plus sulphonylureas versus metformin plus DPP-4 inhibitors: association with major adverse cardiovascular events and all-cause mortality. *Diabetes Obes Metab* 2014;**16**:977–83.
- 55 Morgan CL, Mukherjee J, Jenkins-Jones S, *et al.* Association between first-line monotherapy with sulphonylurea versus metformin and risk of all-cause mortality and cardiovascular events: a retrospective, observational study. *Diabetes Obes Metab* 2014;**16**:957–62.
- 56 Morgan CL, Poole CD, Evans M, *et al.* What next after metformin? A retrospective evaluation of the outcome of second-line, glucose-lowering therapies in people with type 2 diabetes. *J Clin Endocrinol Metab* 2012;**97**:4605–12.
- 57 Schramm TK, Gislason GH, Vaag A, *et al.* Mortality and cardiovascular risk associated with different insulin secretagogues compared with metformin in type 2 diabetes, with or without a previous myocardial infarction: a nationwide study. *Eur Heart J* 2011;**32**:1900–8.
- 58 Phung OJ, Schwartzman E, Allen RW, *et al.* Sulphonylureas and risk of cardiovascular disease: systematic review and meta-analysis. *Diabet Med*

2013;**30**:1160–71.

- 59 European Medicines Agency. Actos (pioglitazone). Eur. public Assess. Rep. <https://www.ema.europa.eu/en/medicines/human/EPAR/actos#overview-section> (accessed 14 Jun 2019).
- 60 Dormandy JA, Charbonnel B, Eckland DJA, *et al.* 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.
- 61 Colmers IN, Bowker SL, Majumdar SR, *et al.* Use of thiazolidinediones and the risk of bladder cancer among people with type 2 diabetes: a meta-analysis. *CMAJ* 2012;**184**:E675-83.
- 62 Mehtälä J, Khanfir H, Bennett D, *et al.* Pioglitazone use and risk of bladder cancer: a systematic literature review and meta-analysis of observational studies. *Diabetol Int* 2019;**10**:24–36.
- 63 Tang H, Shi W, Fu S, *et al.* Pioglitazone and bladder cancer risk: a systematic review and meta-analysis. *Cancer Med* 2018;**7**:1070–80.
- 64 Liao H-W, Saver JL, Wu Y-L, *et al.* Pioglitazone and cardiovascular outcomes in patients with insulin resistance, pre-diabetes and type 2 diabetes: a systematic review and meta-analysis. *BMJ Open* 2017;**7**:e013927.
- 65 Yang J, Vallarino C, Bron M, *et al.* A comparison of all-cause mortality with pioglitazone and insulin in type 2 diabetes: an expanded analysis from a retrospective cohort study. *Curr Med Res Opin* 2014;**30**:2223–31.
- 66 Hippisley-Cox J, Coupland C. Diabetes treatments and risk of heart failure, cardiovascular disease, and all cause mortality: cohort study in primary care. *BMJ* 2016;**354**:i3477.
- 67 Millett ERC, Quint JK, De Stavola BL, *et al.* Improved incidence estimates from linked vs. stand-alone electronic health records. *J Clin Epidemiol* 2016;**75**:66–9.
- 68 Baker R, Tata LJ, Kendrick D, *et al.* Identification of incident poisoning, fracture and burn events using linked primary care, secondary care and mortality data from England: implications for research and surveillance. *Inj*



- Prev* 2016;**22**:59–67.
- 69 Gallagher AM, Williams T, Leufkens HGM, *et al.* The Impact of the Choice of Data Source in Record Linkage Studies Estimating Mortality in Venous Thromboembolism. *PLoS One* 2016;**11**:e0148349.
- 70 What role for metformin in type 1 diabetes? *Drug Ther Bull* 2018;**56**:78–80.
- 71 Bailey CJ. Metformin: historical overview. *Diabetologia* 2017;**60**:1566–76.
- 72 Boggon R, van Staa TP, Chapman M, *et al.* Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiol Drug Saf* 2013;**22**:168–75.
- 73 Pan European Multi-Database Bladder Cancer Risk Characterisation Study (EUPAS3626)).  
<http://www.encepp.eu/encepp/viewResource.htm?id=4510> (accessed 15 Jul 2019).
- 74 Jönsson AK, Schiöler L, Lesén E, *et al.* Influence of refill adherence method when comparing level of adherence for different dosing regimens. *Eur J Clin Pharmacol* 2014;**70**:589–97.
- 75 Vach W, Blettner M. Biased Estimation of the Odds Ratio in Case-Control Studies due to the Use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables. *Am J Epidemiol* 1991;**134**:895–907.
- 76 White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;**29**:2920–31.
- 77 Harrell FJL. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–87.
- 78 Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci* 2010;**25**:1–21.
- 79 Outpatient Data Quality Report - NHS Digital. <https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/outpatient-data-quality-report> (accessed 16 Jul 2019).

- 80 Hospital Outpatient Activity, 2016-17 - NHS Digital.  
<https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/hospital-outpatient-activity-2016-17> (accessed 16 Jul 2019).
- 81 Pearce N, Checkoway H, Kriebel D. Bias in occupational epidemiology studies. *Occup Environ Med* 2007;**64**:562–8.
- 82 The National Casemix Office. HRG4 Companion. 2012.[http://www.hscic.gov.uk/media/10438/HRG4Companionv15pdf/pdf/HRG4\\_Companion\\_v1.5.pdf](http://www.hscic.gov.uk/media/10438/HRG4Companionv15pdf/pdf/HRG4_Companion_v1.5.pdf) (accessed 28 Aug 2014).
- 83 Bazelier MT, Eriksson I, de Vries F, *et al*. Data management and data analysis techniques in pharmacoepidemiological studies using a pre-planned multi-database approach: a systematic literature review. *Pharmacoepidemiol Drug Saf* 2015;**24**:897–905.
- 84 Brooke HL, Talbäck M, Hörnblad J, *et al*. The Swedish cause of death register. *Eur J Epidemiol* 2017;**32**:765–73.
- 85 GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015;**385**:117–71.
- 86 Roulson J, Benbow EW, Hasleton PS. Discrepancies between clinical and autopsy diagnosis and the value of post mortem histology; a meta-analysis and review. *Histopathology* 2005;**47**:551–9.
- 87 Cancer Research UK; National Cancer Intelligence. Cancer by deprivation in England: Incidence, 1996-2010, Mortality, 1997-2011. 2014.<https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bladder-cancer/incidence#ref-4> (accessed 30 Jul 2019).
- 88 GP2GP - NHS Digital. <https://digital.nhs.uk/services/gp2gp> (accessed 4 Mar 2020).
- 89 Alvarez-Madrado S, McTaggart S, Nangle C, *et al*. Data Resource Profile: The Scottish National Prescribing Information System (PIS). *Int J Epidemiol*

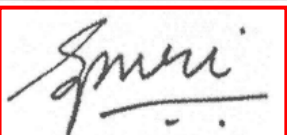
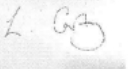



- 2016;**45**:714-715f.
- 90 Oyinlola JO, Campbell J, Kousoulis AA. Is real world evidence influencing practice? A systematic review of CPRD research in NICE guidances. *BMC Health Serv Res* 2016;**16**:299.
- 91 European Medicines Agency (EMA). EMA Regulatory Science to 2025. 2018.[https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection\\_en.pdf](https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection_en.pdf) (accessed 30 Jul 2019).
- 92 Medicines and Healthcare products Regulatory Agency. Corporate Plan 2018-23. 2018.[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/702075/Corporate\\_Plan.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/702075/Corporate_Plan.pdf) (accessed 30 Jul 2019).
- 93 National Institute For Health and Care Excellence (NICE). Consultation on the data and analytics statement of intent. 2019.<https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-guidelines/how-we-develop-nice-guidelines/consultation-data-and-analytics-statement-of-intent> (accessed 30 Jul 2019).
- 94 Strongman H, Williams R, Meeraus W, *et al.* Limitations for health research with restricted data collection from UK primary care. *Pharmacoepidemiol Drug Saf* 2019;;pds.4765.
- 95 Dendup T, Feng X, Clingan S, *et al.* Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. *Int J Environ Res Public Health* 2018;**15**.
- 96 Mourby M, Kaye J, Smith H, *et al.* Health Data Linkage for UK Public Interest Research: Key Obstacles and Solutions. *Int J Popul Data Sci* 2019;**4**:9.
- 97 About | HDR UK. <https://www.hdruk.ac.uk/about/> (accessed 3 Mar 2020).

## Appendix A: Statements of contribution signed by co-authors

### Paper 1:

Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *Eur J Epidemiol.* 2019; 34(1): 91-99

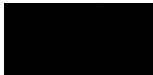
In her role as the observational research lead for linkages, Helen Strongman assessed the potential impact of external changes to information governance and linkage processes on data quality and applicability for research and how new data sources could be used for research. She then worked with the information governance and data, tools and technology teams to update internal processes and research guidance for use of linked data. Helen proposed and promoted the writing of this manuscript, contributed substantially to the scope and structure, and made critical revisions to the content.

I agree that Helen Strongman made the aforementioned contribution to this paper		
Name	Signature	Date
Shivani Padmanabhan		02/05/2019
Lucy Carty		01/05/2019
Ellen Cameron		02/05/2019
Rebecca Ghosh		02/10/2019
Rachael Williams		02/10/2019

Paper 2:

Comparison of diabetes-associated secondary healthcare utilization between alternative oral antihyperglycaemic dual therapy combinations with metformin in patients with type 2 diabetes: An observational cohort study. *Diabetes, Obesity and Metabolism*, 2015; 17(6): 573-80

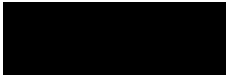
Helen Strongman designed the study (protocol and statistical analysis plan), developed the data management programmes, performed the statistical analysis, and drafted and revised the manuscript.

I agree that Helen Strongman made the aforementioned contribution to this paper		
Name	Signature	Date
Kalpana D'Oca		
Haya Langerman		03.07.2019
Romita Das		

Paper 2:

Comparison of diabetes-associated secondary healthcare utilization between alternative oral antihyperglycaemic dual therapy combinations with metformin in patients with type 2 diabetes: An observational cohort study. *Diabetes, Obesity and Metabolism*, 2015; 17(6): 573-80

Helen Strongman designed the study (protocol and statistical analysis plan), developed the data management programmes, performed the statistical analysis, and drafted and revised the manuscript.

I agree that Helen Strongman made the aforementioned contribution to this paper		
Name	Signature	Date
Kalpana D'Oca		02/07/2019
Haya Langerman		
Romita Das		

Romita Das declined to sign this statement as she stated that she does not own or have any rights over this analysis as she is no longer an employee of MSD.

**Paper 3**

Pioglitazone use and risk of bladder cancer in patients with type 2 diabetes: retrospective cohort study using datasets from four European countries. *BMJ (Clinical Research Ed.)*, 2016; 354:i3903.

This post-authorisation safety study was completed by a consortium of researchers from four European countries. Researchers in each country developed analytical datasets from the raw data and the Finnish team completed the pooled analysis. Helen Strongman programmed the UK analytical datasets using CPRD primary care and linked data sources, and ran statistical analyses as a quality assurance step for comparison with the pooled analyses. Helen also made substantial contributions to protocol design, interpretation of results and manuscript review.

**Paper 4**





Pioglitazone and risk of mortality in patients with type 2 diabetes: results from a European multidatabase cohort study. *BMJ Open Diabetes Research & Care*, 2017; 5(1), e000364.

Helen Strongman led discussions about interpretation of the findings for this analysis and took primary responsibility for drafting and revising the manuscript following comments from co-investigators and peer-reviewers. Responsibility for study design, dataset generation and statistical analysis was the same as for the primary outcome (bladder cancer, described above).

**Paper 5**

Pioglitazone and cause-specific risk of mortality in patients with type 2 diabetes: extended analysis from a European multidatabase cohort study. *BMJ Open Diabetes Res Care*, 2018; 6(1):e000481.

Helen Strongman led discussions about interpretation of the findings for this analysis and took primary responsibility for drafting and revising the manuscript following comments from co-investigators and peer-reviewers. Responsibility for study design, dataset generation and statistical analysis was the same as for the primary outcome (bladder cancer, described above).

I agree that Helen Strongman made the aforementioned contribution to these three papers		
Name	Signature	Date
Pasi Korhonen		15/05/19
Edith Heintjes		11/06/19
Rachael Williams		14/05/19
Fabian Hoti		15/05/19

### Paper 3

Pioglitazone use and risk of bladder cancer in patients with type 2 diabetes: retrospective cohort study using datasets from four European countries. *BMJ (Clinical Research Ed.)*, 2016; 354: i3903.

This post-authorisation safety study was completed by a consortium of researchers from four European countries. Researchers in each country developed analytical datasets from the raw data and the Finnish team completed the pooled analysis. Helen Strongman programmed the UK analytical datasets using CPRD primary care and linked data sources, and ran statistical analyses as a quality assurance step for comparison with the pooled analyses. Helen also made substantial contributions to protocol design, interpretation of results and manuscript review.

### Paper 4



Pioglitazone and risk of mortality in patients with type 2 diabetes: results from a European multidatabase cohort study. *BMJ Open Diabetes Research & Care*, 2017; 5(1), e000364.

Helen Strongman led discussions about interpretation of the findings for this analysis and took primary responsibility for drafting and revising the manuscript following comments from co-investigators and peer-reviewers. Responsibility for study design, dataset generation and statistical analysis was the same as for the primary outcome (bladder cancer, described above).

### Paper 5

Pioglitazone and cause-specific risk of mortality in patients with type 2 diabetes: extended analysis from a European multidatabase cohort study. *BMJ Open Diabetes Res Care*, 2018; 6(1):e000481.

Helen Strongman led discussions about interpretation of the findings for this analysis and took primary responsibility for drafting and revising the manuscript following comments from co-investigators and peer-reviewers. Responsibility for study design, dataset generation and statistical analysis was the same as for the primary outcome (bladder cancer, described above).

I agree that Helen Strongman made the aforementioned contribution to these three papers		
Name	Signature	Date
Solomon Christopher		05 June, 2019
Maila Majak		15 June 2019
Leanne Kool-Houweling (papers 3 and 4 only)		

Leanne Kool-Houweling has changed jobs and I have not been able to find an email address for her or contact her on LinkedIn.

**Paper 3**

Pioglitazone use and risk of bladder cancer in patients with type-2 diabetes: retrospective cohort study using datasets from four European countries. *BMJ (Clinical Research Ed.)*, 2016; 354: i3903.

This post-authorisation safety study was completed by a consortium of researchers from four European countries. Researchers in each country developed analytical datasets from the raw data and the Finnish team completed the pooled analysis. Helen Strongman programmed the UK analytical datasets using CPRD primary care and linked data sources, and ran statistical analyses as a quality assurance step for comparison with the pooled analyses. Helen also made substantial contributions to protocol design, interpretation of results and manuscript review.

**Paper 4**

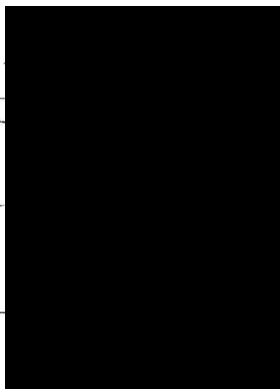
Pioglitazone and risk of mortality in patients with type 2 diabetes: results from a European multidatabase cohort study. *BMJ Open Diabetes Research & Care*, 2017; 5(1), e000364.

Helen Strongman led discussions about interpretation of the findings for this analysis and took primary responsibility for drafting and revising the manuscript following comments from co-investigators and peer-reviewers. Responsibility for study design, dataset generation and statistical analysis was the same as for the primary outcome (bladder cancer, described above).

**Paper 5**

Pioglitazone and cause-specific risk of mortality in patients with type 2 diabetes: extended analysis from a European multidatabase cohort study. *BMJ Open Diabetes Res Care*, 2018; 6(1):e000481.

Helen Strongman led discussions about interpretation of the findings for this analysis and took primary responsibility for drafting and revising the manuscript following comments from co-investigators and peer-reviewers. Responsibility for study design, dataset generation and statistical analysis was the same as for the primary outcome (bladder cancer, described above).

I agree that Helen Strongman made the aforementioned contribution to these three papers		
Name	Signature	Date
Marie Linder		16/5 - 2019
Paul Dolin (papers 3 and 4 only)		11 June 2019
Shahram Bahmanyar		16 / 5 - 2019
Dimitri Bennett (paper 5 only)		May 16, 2019



## Appendix B: Bibliography of works published by the candidate

Rivera DR, Gokhale MN, Reynolds MW, Andrews EB, Chun D, Haynes K, Jonsson-Funk ML, Lynch KE, Lund JL, **Strongman H**, Bhullar H, Raman SR. Linking electronic health data in pharmacoepidemiology: Appropriateness and feasibility. *Pharmacoepidemiol Drug Saf.* 2020 Jan;29(1):18-29

**Strongman H**, Gadd S, Matthews A, Mansfield KE, Stanway S, Lyon AR, Dos-Santos-Silva I, Smeeth L, Bhaskaran K. Medium and long-term risks of specific cardiovascular diseases in survivors of 20 adult cancers: a population-based cohort study using multiple linked UK electronic health records databases. *Lancet.* 2019;394(10203):1041-1054

Carreira H, Williams R, **Strongman H**, Bhaskaran K. Identification of mental health and quality of life outcomes in primary care databases in the UK: a systematic review. *BMJ Open.* 2019; 9(7):e029227

**Strongman H**, Williams R, Meeraus W, et al. Limitations for health research with restricted data collection from UK primary care. *Pharmacoepidemiol Drug Saf* 2019; 28(6):777-787

**Strongman H**, Brown A, Smeeth L, Bhaskaran K. Body mass index and Hodgkin's lymphoma: UK population-based cohort study of 5.8 million individuals. *Br J Cancer.* 2019; 120(7):768-770

Chidwick K, **Strongman H**, Matthews A, Stanway S, Lyon AR, Smeeth L, Bhaskaran K. Statin use in cancer survivors versus the general population: cohort study using primary care data from the UK clinical practice research datalink. *BMC Cancer.* 2018;18(1):1018.

Matthews A, Stanway S, Farmer RE, **Strongman H**, Thomas S, Lyon AR, Smeeth L, Bhaskaran K. Long term adjuvant endocrine therapy and risk of cardiovascular disease in female breast cancer survivors: systematic review. *BMJ.* 2018 ;363:k3845.

Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, **Strongman H**. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *Eur J Epidemiol*. 2018;34(1):91-99.

**Strongman H**, Kausar I, Maher TM. Incidence, Prevalence, and Survival of Patients with Idiopathic Pulmonary Fibrosis in the UK. *Adv Ther*. 2018; 35(5): 724-736

**Strongman H.**, Christopher S., Majak M., Williams R., Bahmanyar S., Linder M., Heintjes EM., Bennett D., Korhonen P., Hoti F. Pioglitazone and cause-specific risk of mortality in patients with type 2 diabetes: extended analysis from a European multidatabase cohort study. *BMJ Open Diabetes Res Care*, 2018; 6(1):e000481.

**Strongman, H.**, Korhonen, P., Williams, R., Bahmanyar, S., Hoti, F., Christopher, S., Majak, M., Kool-Houweling, L., Linder M., Dolin, P., Heintjes, E. M. 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.

Korhonen, P., Heintjes, E. M., Williams, R., Hoti, F., Christopher, S., Majak, M., Kool-Houweling, L., **Strongman, H.**, Linder, M., Dolin, P., Bahmanyar, S. Pioglitazone use and risk of bladder cancer in patients with type 2 diabetes: retrospective cohort study using datasets from four European countries. *BMJ (Clinical Research Ed.)*, 2016; 354: i3903.

**Strongman, H.**, D'Oca, K., Langerman, H., & Das, R. Comparison of diabetes-associated secondary healthcare utilization between alternative oral antihyperglycaemic dual therapy combinations with metformin in patients with type 2 diabetes: An observational cohort study. *Diabetes, Obesity and Metabolism*, 2015; 17(6): 573-80.