

LONDON  
SCHOOL of  
HYGIENE  
& TROPICAL  
MEDICINE



**Methodological issues in electronic healthcare database studies of  
drug cancer associations: identification of cancer, and drivers of  
discrepant results**

**Michael Etrata Rañopa**

**Thesis submitted in accordance with the requirements for the degree  
of Doctor of Philosophy**

**September 2015**

**Department of Non-communicable Disease Epidemiology**

**Faculty of Epidemiology and Population Health**

**London School of Hygiene & Tropical Medicine**

**University of London**

*Funded by the Clinical Practice Research Datalink, as part of  
Pharmacoepidemiological Research on Outcomes of Therapeutics*

## **Declaration**

*I, Michael Etrata Rañopa, confirm that the work presented in this thesis is my own.*

*Where information has been derived from other sources, I confirm that this has been indicated in the thesis.*



Michael Etrata Rañopa

September 2015

## Table of Contents

Declaration .....	2
Acknowledgements .....	9
Abstract .....	10
List of Tables .....	11
List of Figures.....	14
Abbreviations.....	16
1 Background .....	17
1.1 Introduction.....	17
1.2 Safety assessment of new medicines .....	17
1.2.1 Phases I-III: pre-licensing .....	17
1.2.2 Phase IV: post marketing surveillance .....	18
1.3 Bias.....	19
1.4 Electronic health records .....	20
1.4.1 UK primary care databases .....	20
1.4.2 US claims databases.....	21
1.4.3 Scandinavian registries .....	22
1.5 Discrepant results: observational drug safety studies .....	23
1.5.1 Example of conflicting findings: statin use and the risk of cancer .....	25
1.5.2 Statin use and the risk of cancer .....	27
1.6 Rationale for research.....	29
1.7 Aims .....	30
1.8 Outline of thesis.....	30
2 The identification of incident cancers in UK primary care databases: a systematic review .....	32
2.1 Introduction.....	32
2.2 Methods.....	32
2.2.1 Databases and Sources .....	32
2.2.2 Search Keywords and Terms .....	32
2.2.3 Inclusion and Exclusion Criteria.....	33
2.2.4 Procedure .....	33
2.3 Results.....	35

2.3.1	Databases, Cancer Site, and Study Type .....	35
2.3.2	Study Code list Creation, Availability, and Comparison .....	35
2.3.3	Identification and Validation of Cancers.....	39
2.4	Discussion.....	46
2.4.1	Overview .....	46
2.4.2	Accessibility of code lists.....	46
2.4.3	Variation in case definitions and code lists.....	47
2.4.4	External validation of cancer cases .....	48
2.4.5	Comparison of incidence rates.....	48
2.4.6	Limitations of this review.....	49
2.4.7	Importance and implications .....	50
2.4.8	Updated review studies .....	51
2.5	Conclusion .....	51
2.6	Addition of lung cancer to cancer outcomes of interest: rationale .....	52
2.7	Summary .....	52
3	Systematic review of conflicting findings in observational studies utilising electronic patient records to investigate statin use and cancer risk .....	54
3.1	Introduction.....	54
3.2	Aims .....	54
3.3	Methods.....	54
3.3.1	Databases and sources .....	54
3.3.2	Search keywords and terms .....	54
3.3.3	Inclusion and exclusion criteria .....	55
3.3.4	Procedure .....	58
3.3.5	Bias selection .....	58
3.3.6	Meta-analysis .....	59
3.4	Results.....	60
3.4.1	Overview .....	60
3.4.2	Methodological considerations.....	61
3.4.3	Detailed consideration of site-specific associations.....	71
3.4.4	Assessment of bias .....	82
3.4.5	Meta-analysis .....	86
3.5	Discussion.....	91

3.5.1	Overview .....	91
3.5.2	Meta-analysis .....	91
3.5.3	Assessment of bias .....	92
3.5.4	Updated review studies .....	98
3.5.5	Future research .....	99
3.6	Conclusion .....	99
3.7	Summary .....	100
4	Data Sources and Methods .....	101
4.1	Introduction.....	101
4.2	Data Sources.....	101
4.2.1	The Clinical Practice Research Datalink (CPRD) .....	101
4.2.2	Cancer diagnostic groups .....	103
4.2.3	Cancer case definitions .....	106
4.2.4	Externally linked data sources.....	108
4.2.5	ONS Mortality (Death Registry).....	109
4.2.6	Linkage to the CPRD: cancer and death registry .....	110
5	Validity of cancer diagnosis in the CPRD: comparison of observed and expected cancer incidence rates and concordance with national cancer registrations .....	112
5.1	Introduction.....	112
5.2	Objectives.....	112
5.3	Methods .....	112
5.3.1	Patients .....	112
5.3.2	Outcomes .....	113
5.3.3	Follow-up time .....	113
5.3.4	CPRD diagnostic groups and case definitions .....	113
5.3.5	Concordance of recorded cancer diagnoses between primary care, linked cancer registry, and death registry data .....	114
5.3.6	Linked ONS death certificates and cancer registry incidence rates ..	115
5.3.7	Statistical analysis.....	115
5.4	Results.....	119
5.4.1	Cohort .....	119
5.4.2	Potential cases and evidence of diagnosis.....	119

5.4.3	Comparison of incidence rates using CPRD-only case definitions vs ONS published rates .....	123
5.4.4	CPRD linkage to the NCDR and ONS mortality .....	128
5.5	Discussion .....	135
5.5.1	Overview .....	135
5.5.2	Comparison of cancer incidence: CPRD vs ONS .....	135
5.5.3	Alternative case definitions.....	137
5.5.4	Linkage to cancer registry and death registrations .....	139
5.5.5	Cancer Type .....	142
5.5.6	Limitations .....	142
5.5.7	Future Research.....	143
5.6	Conclusion .....	145
5.7	Summary .....	146
6	Systematic evaluation of the impact of potential methodological drivers of discrepant results in a pharmacoepidemiological study of statin use and cancer risk	147
6.1	Introduction.....	147
6.2	Objective .....	147
6.3	Methods .....	147
6.3.1	Primary methodological outcome measure: assessment of potential drivers of discrepant results .....	147
6.3.2	Outcomes .....	149
6.3.3	Treatment groups .....	149
6.3.4	Study design .....	151
6.3.5	Statistical analysis .....	155
6.3.6	Impact of potential drivers of discrepant results .....	157
6.4	Results: Descriptive Analysis .....	174
6.4.1	<i>New statin</i> and non-user cohort .....	174
6.4.2	<i>New statin</i> users vs new glaucoma medication users .....	177
6.4.3	Ever statin user matched cohort .....	178
6.4.4	Nested case-control design.....	179
6.5	Results: Impact of bias .....	182
6.5.1	Overview .....	182

6.5.2	Immortal time bias .....	182
6.5.3	Protopathic bias.....	183
6.5.4	Prevalent user bias .....	186
6.5.5	Healthy user bias .....	186
6.5.6	Time-window bias.....	187
6.5.7	Sensitivity analyses: weighting, missing data, and censoring of treatment change .....	191
6.6	Results: Impact of alternative outcome definitions .....	192
6.6.1	Case definitions .....	192
6.6.2	Linkage to the cancer registry .....	192
6.7	Discussion.....	196
6.7.1	Overview .....	196
6.7.2	Impact of bias on the statin-cancer association.....	196
6.7.3	Sensitivity analysis .....	200
6.7.4	Impact of outcome definition on the statin-cancer association.....	201
6.7.5	Residual bias in corrected analyses .....	202
6.7.6	Limitations.....	203
6.7.7	Future Studies .....	204
6.7.8	Conclusion .....	204
6.8	Summary .....	205
7	Thesis summary and conclusions.....	207
7.1	Introduction.....	207
7.2	Summary of research undertaken .....	207
7.3	Validity of cancer diagnosis in the CPRD (Chapter 5) .....	208
7.3.1	Summary of main findings from Chapter 5.....	208
7.3.2	Validity of recorded cancer diagnoses in the context of previous research .....	210
7.3.3	Strengths of this study .....	211
7.3.4	Limitations of the study .....	212
7.4	Systematic evaluation of the impact of potential methodological drivers of discrepant results in a pharmacoepidemiological study of statin use and cancer risk (Chapter 6) .....	213
7.4.1	Summary of main findings from Chapter 6.....	213

7.4.2	Impact of potential drivers of conflicting results in the context of previous research.....	214
7.4.3	Strengths of the study.....	218
7.4.4	Limitations of the study .....	219
7.4.5	Bias analyses considerations for application to other exemplars of pharmacoepidemiological research.....	219
7.5	Implications of research undertaken.....	221
7.6	Future research .....	222
7.7	Conclusions.....	223
	References.....	225
8	Appendix A - Supplementary materials for Chapter 2.....	243
	Appendix A.1: Chapter 2 systematic review published in the Journal of Pharmacoepidemiology and Drug Safety.....	279
9	Appendix B: Supplementary tables for Chapter 3 .....	287
10	Appendix C: Supplementary tables for Chapter 4 .....	293
	Appendix C.1: ISAC and LSHTM Ethics approval details .....	293
11	Appendix D: Supplementary materials for Chapter 6.....	306



## **Acknowledgements**

First and foremost I would like to thank my supervisors, Dr Krishnan Bhaskaran and Dr Ian Douglas who have been extremely patient, supportive, motivational, and encouraging throughout this project. I am very grateful!

I am especially thankful for the advice and guidance from my advisory committee members, Professor Liam Smeeth and Professor Tjeerd van Staa.

Lastly, I would like to express my deepest love and gratitude to my parents and to my fiancée Laura who has been my clear head and light during the darkest hours of this project.

## Abstract

**Background:** There have been a number of conflicting findings from epidemiological studies investigating the association of drug use and cancer risk. Methodological issues such as biased study designs and differences in case identification have been postulated as potential reasons for differing results. However, the impact of these methodological variants is unclear.

**Aims:** The principal aims of this thesis were to develop and validate case definitions that identified incident cancer diagnosis in the Clinical Practice Research Datalink (CPRD), and to measure and compare the impact of several potential drivers of conflicting findings within a practical setting.

**Methods:** Firstly, for breast, colorectal, lung, and prostate cancer, two sets of incidence rates were estimated and compared to national estimates: (i) based on cancers identified in the CPRD; and (ii) estimates from the CPRD incorporating linked cancer registry data. Secondly, the statin-cancer association was investigated as an exemplar, and several potential drivers of conflicting findings were examined including study bias, case definitions, and data linkage. Study bias included immortal time, protopathic, prevalent user, healthy user, and time-window bias.

**Results:** Cancer incidence rates based on the CPRD alone were lower compared to national estimates across all cancer types. Compared to national estimates, incidence rates incorporating linked cancer registry data were similar for colorectal and lung cancer, but higher for breast and prostate cancer. Of the seven potential drivers of discrepant results in the example study of statins and cancer, only time-window bias yielded substantial and consistent biased effects, with bias towards a protective association and corrected analyses yielding a null association. Immortal time, protopathic, prevalent user, and healthy user bias had minimal impact on the estimated association between statin use and cancer risk.

**Conclusions:** CPRD cancer incidence rates were lower compared to national estimates. Incorporating linked cancer registry data, breast and prostate cancer incidence rates were higher than expected, implying that a proportion of the cancer cases identified in the CPRD were either false-positive cases or not registered nationally. A number of common design flaws and decisions were postulated as drivers of discrepant results. However, in practical study settings these flaws and differences did not uniformly lead to large changes in the estimated association of statin use and cancer risk.

## List of Tables

Table 2.1: Code list availability, questionnaire replies, and comparison of lists received, by cancer and study type .....	38
Table 2.2: Criteria used to identify, validate, and exclude potential cancer cases by cancer and study type .....	41
Table 2.3: External validation of potential cases by cancer type.....	42
Table 2.4: Comparison of incidence rates by cancer type .....	45
Table 3.1: Description of biases assessed.....	56
Table 3.2: Frequency (%) of findings by cancer type .....	61
Table 3.3: Statin use associated with cancer risk - study details .....	62
Table 3.4: Summary of findings and biases - risk of breast, prostate, and colorectal and lung cancer associated with statin use .....	67
Table 3.5: Association between duration of statin use and cancer risk – secondary analysis results .....	88
Table 4.1: Case definitions .....	107
Table 5.1: Distribution of age (at cohort entry) and sex for the random sample of 2 million patients from the CPRD .....	117
Table 5.2: Comparison of CPRD age-standardised incidence rates (2000-2010) compared to published ONS age-standardised rates, by cancer type and case definition .....	122
Table 5.3: Cases identified in the cancer registry alone .....	131
Table 6.1: Demographics for the matched <i>new statin</i> cohort and nested case-control design (risk-set sampling) .....	175
Table 6.2: Immortal time bias relative risk estimates, $\Delta\beta$ estimates and corresponding 95% confidence intervals.....	184
Table 6.3: Protopathic bias relative risk estimates, $\Delta\beta$ estimates and corresponding 95% confidence intervals.....	185
Table 6.4: Prevalent user bias relative risk estimates, $\Delta\beta$ estimates and corresponding 95% confidence intervals.....	188
Table 6.5: Healthy user bias relative risk estimates, $\Delta\beta$ estimates and corresponding 95% confidence intervals.....	189

<b>Table 6.6: Time-window bias relative risk estimates, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>190</b>
<b>Table 6.7: Case definitions relative risk estimates, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>194</b>
<b>Table 6.8: Linked data relative risk estimates, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>195</b>
<b>Table 8.1: Summary description table of included studies .....</b>	<b>245</b>
<b>Table 8.2: Frequency of studies that included specific cancer related codes .....</b>	<b>261</b>
<b>Table 9.1: Detailed summary of bias assessment by study .....</b>	<b>287</b>
<b>Table 9.2: Updated review studies: Statin use associated with cancer risk - study details .....</b>	<b>290</b>
<b>Table 9.3: Summary of Findings and Biases - Risk of breast, prostate, and colorectal and lung cancer associated with Statin Use.....</b>	<b>292</b>
<b>Table 10.1: Breast cancer code list .....</b>	<b>294</b>
<b>Table 10.2: Colorectal cancer code list .....</b>	<b>297</b>
<b>Table 10.3: Prostate cancer code list.....</b>	<b>300</b>
<b>Table 10.4: Lung Cancer Code list .....</b>	<b>301</b>
<b>Table 10.5: Malignant Neoplasm (Site Unknown) codelist.....</b>	<b>304</b>
<b>Table 10.6: Supporting evidence of diagnosis .....</b>	<b>305</b>
<b>Table 11.1: Demographics for the cohort of any statin users; glaucoma cohort.....</b>	<b>306</b>
<b>Table 11.2: Demographics for the Case-control design (time-independent sampling).....</b>	<b>308</b>
<b>Table 11.3: Immortal time bias weighted relative risk, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>309</b>
<b>Table 11.4: Protopathic bias weighted relative risk, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>310</b>
<b>Table 11.5: Prevalent user bias weighted relative risk, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>311</b>
<b>Table 11.6: Immortal time bias imputed, missing category relative risk estimates, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>312</b>

<b>Table 11.7: Protopathic bias imputed, missing category relative risk estimates, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>313</b>
<b>Table 11.8: Prevalent user bias imputed, missing category relative risk estimates, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>314</b>
<b>Table 11.9: Healthy user bias imputed, missing category relative risk estimates, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>315</b>
<b>Table 11.10: Time-window bias imputed, missing category relative risk estimates, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>316</b>
<b>Table 11.11: Immortal time bias censored relative risk, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>317</b>
<b>Table 11.12: Protopathic bias censored relative risk, <math>\Delta\beta</math> estimates and corresponding 95% confidence intervals.....</b>	<b>318</b>
<b>Table 11.13: Prevalent user bias censored and primary analysis relative risk estimates, corresponding 95% confidence intervals, and percentage difference in risk attributable to prevalent user bias .....</b>	<b>319</b>

## List of Figures

Figure 2.1: Flow diagram of article search, retrieval, and review process; code list availability and questionnaire replies; database, study type, and validation methods. ....	36
Figure 3.1: Diagrammatic illustration of biases examined .....	57
Figure 3.2: Flow diagram of search strategy: inclusion/exclusion criteria.....	59
Figure 3.3: Forest plots of studies examining statin use and breast, colorectal, lung, and prostate cancer risk .....	73
Figure 4.1: CPRD diagnostic algorithm .....	105
Figure 4.2: Flow diagram illustrating sources of cancer registrations [reproduced from Cancer Statistics registrations Series MB1 <sup>180</sup> ] .....	109
Figure 5.1: Flow diagram of cancer cohorts - inclusion and exclusion.....	118
Figure 5.2: Number of potential cases by diagnostic group and corresponding proportion of evidence of diagnosis.....	121
Figure 5.3: Comparison of CPRD and ONS age-standardised incidence rates by calendar year.....	124
Figure 5.4: Comparison of primary care and ONS reported crude age-specific incidence rates (over the whole study period) by age at diagnosis.....	125
Figure 5.5: Flow diagram of eligible CPRD (primary care) patients linked to the cancer registry (NCDR) and death registry (ONS mortality) .....	127
Figure 5.6: Agreement of recorded diagnosis between the CPRD, NCDR, and ONS mortality by cancer type .....	130
Figure 5.7: Primary age-standardised incidence rates incorporating cancer and death registry linked data by cancer type and linkage .....	134
Figure 6.1: Diagrammatic representation of intention to treat analysis .....	152
Figure 6.2: Immortal time bias: biased and corrected designs .....	161
Figure 6.3: Protopathic bias: biased and corrected designs .....	161
Figure 6.4: Prevalent User bias: biased and corrected designs.....	162
Figure 6.5: Time-window bias: biased and corrected designs .....	162
Figure 6.6: Flow diagram of inclusion exclusion of new statin users matched to non-users .....	170

<b>Figure 6.7: Flow diagram of inclusion exclusion of new statin and new glaucoma medication users.....</b>	<b>171</b>
<b>Figure 6.8: Flow diagram of inclusion exclusion of statin users (incident and prevalent) matched to non-users.....</b>	<b>172</b>
<b>Figure 6.9: Flow diagram of the cohort of new statin users and non-users and a case-control design nested within the cohort .....</b>	<b>173</b>
<b>Figure 6.10: Distribution of treatment start date by age and treatment group .....</b>	<b>177</b>
<b>Figure 6.11: Relative risk estimates and corresponding 95% confidence intervals for each bias analysis by cancer type.....</b>	<b>181</b>

## Abbreviations

BNF	British National Formulary
CPRD	Clinical Practice Research Datalink
CI	Confidence Interval
EHR	Electronic Health Records
GP	General Practitioner
HRT	Hormone Replacement Therapy
HR	Hazard Ratio
ICD-9	International Classification of Diseases, 9 <sup>th</sup> Revision
ICD-10	International Classification of Diseases, 10 <sup>th</sup> Revision
IR	Incidence rate
NHS	National Health Service
NCDR	National Cancer Data Repository
ONS	Office for National Statistics
OR	Odds ratio
OC	Oral contraceptive
PPV	Positive Predictive Value
RR	Relative Risk
RCT	Randomised Controlled Trial
THIN	The Health Information Network
UK	United Kingdom
UTS	Up-to-standard



# **1 Background**

## **1.1 Introduction**

In this chapter, observational pharmacoepidemiological studies are first placed in the context of the regulatory stages of drug development and safety assessment. Second, the concept of bias in observational studies is introduced, including a description of the specific biases that will be examined in this thesis. Third, UK primary care databases are introduced with particular focus on methods related to case identification and validation for such databases. Lastly, there is a brief overview of interventional and observational studies that have examined the risk of cancer among patients prescribed statins.

## **1.2 Safety assessment of new medicines**

### **1.2.1 Phases I-III: pre-licensing**

Before a medication can be widely used, it must first be subjected to a series of clinical tests before a license is granted (marketing authorisation). There are three clinical phases in the drug development process prior to market authorisation.<sup>1</sup> Phase I trials assess the medication's safety, tolerability, and pharmacokinetics of escalating doses, typically among a small (<100) number of usually healthy volunteers. Phase II trials test the efficacy of the medication, i.e. how well the medication works at selected doses, as well as continuation of phase I safety assessment in a moderate number of people (several hundred) with the target condition or disease. Phase III studies (randomised controlled trials, RCT) aim to assess the effectiveness of the new medication usually in comparison with the current "gold standard" treatment. Phase III studies are conducted in several hundred to several thousand people over a period that can last for several years.

Phase IV studies (post market surveillance) are conducted once a drug has been approved for use in patients.<sup>1</sup>

### **1.2.2 Phase IV: post marketing surveillance**

Once the medication is approved for use among patients, further monitoring outside clinical trial settings is required. Post-marketing surveillance is generally divided into two main stages: signal detection and signal evaluation.

The importance of post-marketing surveillance is illustrated by the withdrawal of co-proxamol by the Medicines and Healthcare products Regulatory Agency (MHRA), over an increased risk of suicide amongst patients prescribed this medication.<sup>2</sup>

#### **1.2.2.1 Signal detection**

After licensing is approved, safety of medications is predominantly monitored through a system of spontaneous reports. In the UK, the MHRA oversees the Yellow Card Scheme, whereby healthcare professionals and patients themselves report adverse events via a “Yellow Card” form.<sup>3</sup> Once spontaneous reports are collected they are screened for signals and are then evaluated further in terms of causality, frequency, clinical implications, and preventability.

#### **1.2.2.2 Observational drug safety studies**

Once potential signals have been identified, further evaluation is needed to decide whether the medicinal product should be maintained, changed, suspended or withdrawn. Evaluation of signals can be assessed either by conducting further clinical trials or observational (pharmacoepidemiological) studies, for example, cohort or case-control studies. The importance of such observational studies is illustrated by their contribution toward the decision-making process of over a third

of all drugs withdrawn (due to safety concerns) from the European market between 2002-2011.<sup>4</sup>

There has been a surge in observational studies conducted in recent years.<sup>5</sup> This increase has undoubtedly led to significant contributions to existing medical literature. However, cautious interpretation of such findings is needed due to the susceptibility of observational studies to bias.

### **1.3 Bias**

The concept of bias is the lack of internal validity or incorrect assessment of the association between an exposure and outcome which deviates from the true relationship.<sup>6</sup> Biases are often classified into three main groups: (i) selection bias, (ii) information bias, and (iii) confounding. Selection bias is related to study subject recruitment or retention procedures. Information bias is concerned with procedures used to measure the information about study variables and confounding is the distortion caused by other variables related to both exposure and outcome.

Unlike RCTs, pharmacoepidemiological studies are particularly prone to confounding bias due to their observational nature. More specifically, RCTs limit the potential of confounding (observed or unobserved) by randomly allocating subjects to a treatment group, whereby the chance of observing differences between treatment groups is minimised. Epidemiological studies are observational in the sense that no experimental intervention takes place; events that take place are an occurrence of “real world” settings. Although experimental studies may be less susceptible to confounding, they have several limitations with respect to their use in safety studies. First, trials are generally not powered to assess secondary or

relatively uncommon safety outcomes such as cancer. Second, follow-up periods in trials are relatively short due to the cumulative cost of long-running trials which limits their capability of examining outcomes with long latency periods (e.g. cancer). Last, eligibility criteria and drug-dose selection in RCTs often do not reflect the general patient population or drug dose for which the medicine will eventually be used, which limits generalisability.

Bias in medical studies are numerous and well documented, some of which are specific to study-type (e.g. clinical trials, and ecological studies). This thesis concentrates on a subset of biases that have been commonly cited as possible reasons for conflicting results in pharmacoepidemiological studies.

A systematic review of an established drug-cancer association is described in **Chapter 3**, with a detailed focus on specific biases including immortal time, protopathic, healthy user, prevalent user, and time-window bias.

## **1.4 Electronic health records**

Existing data sources containing longitudinal health data can be utilised by observational studies to answer questions about suspected adverse events from medications. The following sections provide a summary of the various sources of routinely collected data that can be utilised for pharmacoepidemiological research.

### **1.4.1 UK primary care databases**

In UK primary care settings, general practitioners (GP) are often the first point of contact and continuing point of care for most patients. The UK National Health Service (NHS) provides universal primary care coverage to most UK residents which extends from birth. GPs provide a wide range of health services, including: advice

about health problems, vaccinations, examinations and treatment, prescriptions for medicines, and referrals to secondary health services.

Currently, there are several databases in the UK that collect data from GP practices, three of the principal ones being: the Clinical Practice Research Datalink (CPRD; formerly known as the General Practice Research Database, or GPRD - throughout the remainder of this thesis, the GPRD will be referred to as the CPRD), QRSESEARCH, and The Health Information Network (THIN). Intricacies of the recording systems used by each database differ slightly; the CPRD and THIN use the Vision IT system, while QRSESEARCH uses EMIS software. To date, primary care databases have been used to conduct observational research covering many broad themes, such as: pharmacoepidemiology, drug utilisation, public health, and health services research.<sup>7</sup>

#### **1.4.2 US claims databases**

The US healthcare system generally consists of three main entities: (i) beneficiaries e.g. patients; (ii) healthcare providers e.g. clinics, and hospitals; and (iii) “payers”, e.g. the US government, private health insurance companies, and patients (self-payers). In general, utilisation of healthcare services by a patient (beneficiary) in the US incurs a cost: a healthcare provider requests payment (a “claim”) which is usually sent to the “payer”. The majority of US residents are either covered by the government or privately insured, mainly as part of a health program offered by their employer.<sup>8</sup>

Examples of government co-ordinated healthcare programmes include Medicaid, Medicare, and the Veterans Affairs program.<sup>8</sup> Eligibility criteria for enrolment in these programmes vary: Medicaid provides medical coverage for low-income

individuals and families without private health insurance. Medicare provides medical insurance to US residents aged 65 years or older, and the Veterans Affairs program covers beneficiaries who have served in the US military. Private insurance companies, such as Kaiser Permanente, offer medical coverage to members. Available health plans are usually employee-sponsored but can also be purchased privately. Medical coverage provided by these organisations varies but at the bare minimum includes hospital, medical and prescription drug costs.

Data from such claims databases have been used in various observational studies examining questions related to drug utilisation, drug safety, and comparative effectiveness.<sup>9-15</sup>

### **1.4.3 Scandinavian registries**

In the Nordic countries (Denmark, Finland, Iceland, Norway, and Sweden), several patient registries are available for epidemiological research. Although the healthcare systems across the Nordic countries are not identical they all share similar characteristics. In all Nordic countries, the national government is largely responsible for the co-ordination and financing of primary and secondary health care.<sup>8</sup> All Nordic residents are provided tax-supported health care by the national health service, a personal identification number is allocated to all citizens at birth (and immigrants) by the respective tax agencies as part of a population register. The population register can be linked unambiguously to the nationwide prescription and disease registers including medical birth, cancer, causes of death, and hospital discharge.<sup>8,16</sup>

Since the 1980s, pharmacies in Nordic countries have computerised their records enabling archiving of dispensed prescriptions. As part of the national public health

insurance, universal coverage with unrestricted access to healthcare services is provided as well as partial or complete reimbursement for the cost of medicines prescribed by a physician.<sup>16</sup> The data collected are determined by country-specific regulations but all include information on dispensed and prescribed prescriptions, clinical diagnoses, and patient demographics together with information from different administrative registries. According to the legislation of each country, no informed consent is required for collection of the prescription data, but individuals may seek information about themselves by request.<sup>16</sup>

### **1.5 Discrepant results: observational drug safety studies**

An increasing number of observational drug safety studies have utilised electronic data sources from routinely collected healthcare data.<sup>17,18</sup> However, conflicting findings between such studies limit their usefulness when assessing the benefit or risks posed by marketed medicinal products. Differing populations might contribute to the discrepancy of study results. Moreover, different data sources collect information for various reasons, which directly influences what is collected from patients. For example, a drug-cancer study conducted in a US claims database may not be able to adjust for smoking status because lifestyle factors are not routinely collected from insurance claims. In contrast, a study utilising a UK primary care database might be able to adjust for the effect of smoking status as a confounding factor on the drug-cancer association.

Conflicting findings from studies conducted in the same data source limit the variability introduced by differing populations. Within the same study population, differing results might arise due to small changes in study conduct and design choices, such as: outcome definitions and ascertainment; exposure definitions, age-

matching, and control sampling. There are a number of recent conflicting findings from pharmacoepidemiological studies conducted in the same data source; for example: statin use and risk of fracture,<sup>19, 20</sup> oral bisphosphonates and cancer risk,<sup>21-23</sup> diabetes medications and cancer risk, and statin use associated with cancer risk.<sup>24-30</sup>

In a study utilising the CPRD, fracture risk was found to be associated with an 88% reduction in statin users compared to non-users (OR= 0.12; 95% CI; 0.04, 0.40).<sup>31</sup> However, another study conducted in the CPRD reported no significant effect of statin use on fracture risk (OR=0.59; 95% CI; 0.31, 1.13). de Vries *et al.*<sup>19</sup> examined design differences between the two studies and described a number of variations that might have contributed to the conflicting findings including: case definitions, age-matching, and time-window of exposure to statin use.

Similarly, studies from different settings have shown conflicting findings when examining the association between the use of diabetes related medications and the risk of cancer. For example, a study set in Germany reported an increased incidence of malignancy among patients taking insulin.<sup>32</sup> In contrast, a UK database study observed no significant effects of insulin on cancer risk.<sup>33</sup> In a commentary by Pocock *et al.*<sup>34</sup> methods employed by the German study may have been subject to selection bias when defining exposure status; partially explaining the increased risk of cancer. Exposure time related biases were also shown to have potentially affected studies showing a decreased risk of cancer associated with metformin use.<sup>35</sup>



## **1.5.1 Example of conflicting findings: statin use and the risk of cancer**

### **1.5.1.1 Statins**

Statins are effective hypolipidemic drugs commonly used to treat hypercholesterolemia and to prevent cardiovascular disease. They are among the most commonly prescribed drugs worldwide. Simvastatin was the first statin to be introduced to the UK market in 1989; the early 1990s saw the release of pravastatin, fluvastatin, and lovastatin in the UK. The use of statins has increased dramatically over the last decade following reports from RCTs of substantial risk reductions in cardiovascular disease and related mortality.<sup>36-41</sup> The ageing population and recent availability of over-the-counter low-dose statin formulations are likely to continue their increased utilisation. Apart from cardiovascular disease, there are increasing questions being raised about possible protective properties against other diseases and conditions including cancer,<sup>42-44</sup> dementia,<sup>45</sup> and fractures.<sup>31</sup>

### **1.5.1.2 Cancer**

Cancer is a major public-health issue worldwide, with approximately 14 million new cases and 8.2 million cancer related deaths worldwide in 2012.<sup>46</sup> Cancer, known medically as a malignant neoplasm, is a broad group of diseases. A defining feature of cancer is the division of cells which grow uncontrollably, forming malignant tumours.<sup>46</sup>

### **1.5.1.3 Risk Factors**

The causes of cancer are diverse, complex, and only partially understood. Known risk factors vary by cancer type and their effects are more pronounced among some cancer types. Many factors affect the risk of cancer, including tobacco use, dietary

factors, certain infections and co-morbidities, exposure to radiation, physical activity, obesity, and environmental factors.<sup>46</sup> For breast cancer, specific risk factors include use of hormone replacement therapy and oral contraceptive use. For colorectal cancer, several risk factors have been established in epidemiological studies such as family history of colorectal cancer, inflammatory bowel disease, smoking, excessive alcohol consumption, and high consumption of red and processed meats. The most important risk factor for lung cancer is tobacco smoking, with evidence suggesting that 90% of lung cancer cases can be attributed to tobacco smoking.<sup>47</sup> Unlike other cancer types mentioned, risk factors for prostate cancer have not been well established. Age, ethnicity, and geography are strongly correlated with the risk of prostate cancer: however, there is no evidence suggesting a link to smoking tobacco. Some studies have suggested a link between prostate cancer and diet as well as obesity, although the overall findings are inconclusive.

#### **1.5.1.4 Diagnosis and screening**

Cancer can be detected in a number of ways, including: the presence of certain signs and symptoms, screening tests, or medical imaging. Once a possible cancer is detected, it is diagnosed by microscopic examination of a tissue sample. In the UK, the NHS offers screening programmes for breast, cervical, and colorectal cancer. Screening for breast cancer involves a mammogram<sup>8</sup> of both breasts; women aged 47-73 years are invited for breast cancer screening every 3 years. Colorectal cancer screening in the UK involves a process called Faecal Occult Blood Testing. Testing kits are sent to eligible patients and processed on return. Varying ages are eligible for colorectal cancer screening depending on region. In England, men and women

aged between 60-69 years are invited for colorectal cancer screening, while in Northern Ireland and Wales the age range is 60-74 years, and in Scotland it is 50-74 years.

#### **1.5.1.5 Treatment**

Each cancer type requires a specific treatment plan which depends on the development (stage and grade) of the cancer at diagnosis. A wide range of treatment types are used, including: surgery, chemotherapy, radiation therapy, and biological therapy. In some cases, active monitoring of the cancer is considered the optimum management strategy.

#### **1.5.2 Statin use and the risk of cancer**

A brief overview of findings from experimental and observational studies examining the risk of cancer associated with statin use is given in the following sections. A detailed systematic review of statin use and cancer risk among studies that have utilised electronic healthcare records is described in **Chapter 3** with a focus on methodological considerations.

##### **1.5.2.1 Early findings**

Early concerns of carcinogenic properties from statins were first raised in animal studies. A review of findings on rodent carcinogenicity of lipid-lowering drugs reported that all statins available in 1994 initiated or promoted cancer in rodents at concentrations equivalent to those commonly prescribed in humans.<sup>48</sup> Later that year, a clinical trial lasting 5 years examining pravastatin for the prevention of cardiovascular disease found an increased risk of breast cancer in subjects randomised to statins (12 cases vs. ref. 1 case;  $p=0.002$ ).<sup>49</sup> In a separate trial, an increased risk of total and gastrointestinal cancer was found.<sup>41</sup> However, no other

large RCTs of statins have demonstrated an altered risk of incident cancer.<sup>40, 50, 51</sup>

Findings from a meta-analysis of 35 RCTs found no association between statin use and cancer risk.<sup>52</sup> Evidence from clinical trials should be considered as part of a drug-safety assessment, but these trials were generally not powered to assess uncommon or delayed safety outcomes such as cancer. Therefore, further investigation from observational studies was required to confirm or refute the limited evidence from trials indicating an association between statin and cancer risk.

### **1.5.2.2 Pharmacoepidemiological studies**

A number of studies have examined breast, colorectal, lung, and prostate cancer risk associated with statin use. Aggregations of results from observational drug safety studies do not lend themselves to support an association between statin use and the risk of cancer.<sup>53</sup> However, there have been reports of increased or reduced risks of cancer associated with statin use from observational studies. A cohort study by Cauley *et al.*<sup>54</sup> found hydrophobic (simvastatin, lovastatin, and fluvastatin) statin use to be associated with an 18% (HR=0.82; 95% CI 0.70-0.97) risk reduction of breast cancer. Coogan *et al.*<sup>55</sup> conducted a case-control study reporting a non-significant increased risk of breast cancer associated with statin use (OR=1.5; 95% CI, 1.0-2.3). In contrast, Smeeth *et al.*<sup>56</sup> observed a null association between statin use and breast cancer risk (HR=1.17; 95% CI 0.95-1.43).

A meta-analysis conducted by Bonovas *et al.*<sup>57</sup> observed a null association between statin use and the risk of colorectal cancer in three cohort studies (rate ratio = 0.96; 95% CI, 0.84-1.11). In contrast, Poynter *et al.*<sup>58</sup> conducted a case-control study based in Israel and found a reduced risk of colorectal cancer among statin users

(adjusted OR=0.53; 95% CI, 0.38-0.74). This decreased risk was also reported by Bonovas *et al.*<sup>57</sup> in a meta-analysis of nine case-control studies (RR = 0.91; 95% CI, 0.87-0.96).

Several case-control studies reported of no strong associations between statin use and lung cancer risk.<sup>59-63</sup> A cohort study conducted in Canada<sup>64</sup> reported a borderline increased risk of colorectal cancer associated with statin use (rate ratio=1.13; 95% CI; 1.02, 1.25). Findings from studies that utilised patient data from veteran populations in the US have found strong reduced risk estimates of lung cancer ranging from 30-40% risk reduction.<sup>26, 27</sup> In contrast, Friedman *et al.* observed an increased incidence of lung cancer among women prescribed statins (HR=1.16; 95% CI 1.06-1.28).

Two case-control studies utilising data from veteran populations in the US reported a reduced risk of prostate cancer, ranging from 54-65% reduced risk.<sup>65, 66</sup> There have been occasions where an increased risk of prostate cancer have been reported including: a cohort study by Haukka *et al.*<sup>67</sup>, a case-control study conducted in Taiwan,<sup>68</sup> and a non-significant 30% increase observed by Kaye *et al.*<sup>62</sup> in the CPRD

## **1.6 Rationale for research**

The emergence of electronic patient records has seen a wealth of pharmacoepidemiological studies investigating various drug-cancer associations.<sup>5, 7</sup> However, there have been inconsistent findings from some of these studies<sup>19, 21</sup> which have led to questions about possible explanations of such deviations in observed findings. Various design flaws and study decisions have been postulated as drivers of these conflicting findings.<sup>35, 69-71</sup> However, the impact of such methodological variants in a practical setting is unclear. Based on this rationale, the

impact of several commonly cited biases and alternative methods of case identification on a well-established association, namely the risk of cancer among patients prescribed statins, will be examined in this thesis.

## **1.7 Aims**

1. To review the methods utilised by current and past studies that have identified breast, colorectal, and prostate cancer cases from UK primary care databases.
2. To review the literature to date regarding the effects of statin use and the risk of breast, colorectal, lung, and prostate cancer.
3. To develop case definitions to identify breast, colorectal, lung, and prostate cancers using primary care data, and to validate these definitions by comparing primary care incidence rates to published national rates based on cancer registrations, and by assessing the agreement of recorded cancer diagnoses between primary care data and linked cancer registry data.
4. To measure and compare the impact of several potentially biased study designs and case identification methods on the estimated association between statin use and cancer risk.

## **1.8 Outline of thesis**

**Chapter 2** describes a systematic review of breast, colorectal, and prostate cancer case identification methods from studies that have utilised UK primary care databases. From this review there was evidence suggesting fatality may influence case ascertainment in primary care data. In order to investigate this hypothesis lung cancer was added to the existing three cancer types examined in this thesis.

**Chapters 3** describes a systematic review of studies that have utilised electronic patient records to examine the risk of breast, colorectal, lung, and prostate cancer among statin users. In particular, methodological aspects of each study will be assessed.

**Chapter 4** describes the main data sources and an overview of the analytical methods applied to this thesis, including the main case definitions that were developed for breast, colorectal, lung, and prostate cancers.

**Chapters 5 and 6** present the main analyses of this thesis; in each of these chapters the specific methods are detailed and the corresponding results are presented and discussed.

**Chapter 5** presents the first main analysis of this thesis, which measures agreement of recorded diagnosis from primary care (CPRD) compared to the cancer registry. In addition, this chapter also includes two sets of estimated incidence rates which are compared to national rates based on cancer registrations. The first set includes estimated incidence rates from primary care. The second set includes primary care incidence rates that incorporate linked cancer registry data.

**Chapter 6** presents the second main analysis of this thesis, which measures and compares the impact of potential drivers of discrepant results within the context of the statin-cancer association.

**Chapter 7** summarises the main findings and discussion points from the body of this thesis.

## **2 The identification of incident cancers in UK primary care databases: a systematic review**

### **2.1 Introduction**

This chapter reviews case ascertainment methods implemented by past observational studies that have utilised UK primary care databases to investigate incident cancer outcomes of the breast, colorectum, and prostate. A version of this chapter was published in the Journal of Pharmacoepidemiology and Drug Safety, the full paper is provided in **Appendix A.1**.

### **2.2 Methods**

#### **2.2.1 Databases and Sources**

MEDLINE and EMBASE were searched between January, 1980 – April, 2013 using MeSH terms. Reference lists of relevant studies were also screened for publications that may have been missed by the initial database search. Bibliographies of the CPRD, THIN, QRESEARCH, and the Boston Collaborative Drug Surveillance Program were also screened to identify additional articles that may have been missed by the initial search.<sup>17, 18, 72, 73 74-76</sup>

#### **2.2.2 Search Keywords and Terms**

The search of MEDLINE (April 8, 2013) included exploded key terms to identify publications that utilised a UK primary care database and examined incident cancer as an outcome of interest. For EMBASE, which does not use the MeSH classification system, the nearest equivalent search terms from the EMBASE indexing system were used.



### **MEDLINE MeSH terms:**

The following MeSH keywords were used in the primary search:

[Malignant or Cancer or Neoplasm (plus all sub-terms in the MeSH tree)]  
and

[ [GPRD; CPRD; THIN; QRESEARCH; and DIN-LINK (and exploded synonyms)]  
or [Database (plus all sub-terms in the classification tree)] ]

### **2.2.3 Inclusion and Exclusion Criteria**

A publication was considered for initial inclusion when incident cancers of the breast, colorectum, or prostate were included as primary or secondary outcomes, and a UK primary care database was utilised as a data source. Studies with a main outcome of prevalent, recurring, or metastatic cancer were excluded. Articles presented as conference abstracts, review articles, or letters to the editor were also excluded.

### **2.2.4 Procedure**

Titles and abstracts were initially screened; full-text versions were then obtained and examined to determine whether they met inclusion criteria. Data were extracted from each manuscript and included first author, year of publication, study type (e.g. drug safety, epidemiological, or incidence), database(s), cancer outcome(s) of interest, methods used to create code lists (as reported in the paper, e.g. methods section or supplementary material), case definitions, validation methods and results.

For each study, an electronic copy of the study code list was requested and the first author was sent a questionnaire which included specific questions on the development of their code list(s). Details of the questionnaire are given in **Appendix A, Questionnaire A.1**. Three emails were sent to authors. First, the corresponding

author was contacted; if no response was received after 3 weeks then a reminder email was sent to the same author and additionally to the first or last author (if different). A final reminder was sent after a further 3 weeks if necessary. If an error reply was received stating that the email address had expired, a search for a current email address in more recent publications and through an internet search engine was conducted.

Medical codes were classified into eight groups: malignant neoplasms, in-situ tumours, malignant morphology; secondary or history of cancer, borderline (uncertain whether malignant or benign), suspected (suspected cancer, abnormal screening test, or fast track referral), benign tumours, and non-cancerous codes (procedure, or condition that was not related to a direct malignant neoplasm diagnosis). Codes were stratified by cancer site and study type. The ICD-10 dictionary and medical references were used to aid in the classification of OXMIS and Read codes.<sup>76-79</sup> All codes were reviewed and classified by Krishnan Bhaskaran (KB), Michael Rañopa (MR), and Liam Smeeth (LS); any disagreements were reviewed again until resolved.

All studies in the review were published after the release of the 5-byte Read or Read version 2 dictionary, therefore study code lists were based on the same broad dictionary version. However, codes are continually added to the dictionary over time (though never removed). To assess whether variation in study code lists might have been driven by such changes over time, a full list of code additions (updates were documented every 6 months from 1991-2013) from the NHS Health and Social Care Information Centre (HSCIC) was obtained. This HSCIC code list was used this to

identify codes added during the time period over which the studies were conducted (which was assumed to be in the 2-year period prior to year of publication).

## **2.3 Results**

### **2.3.1 Databases, Cancer Site, and Study Type**

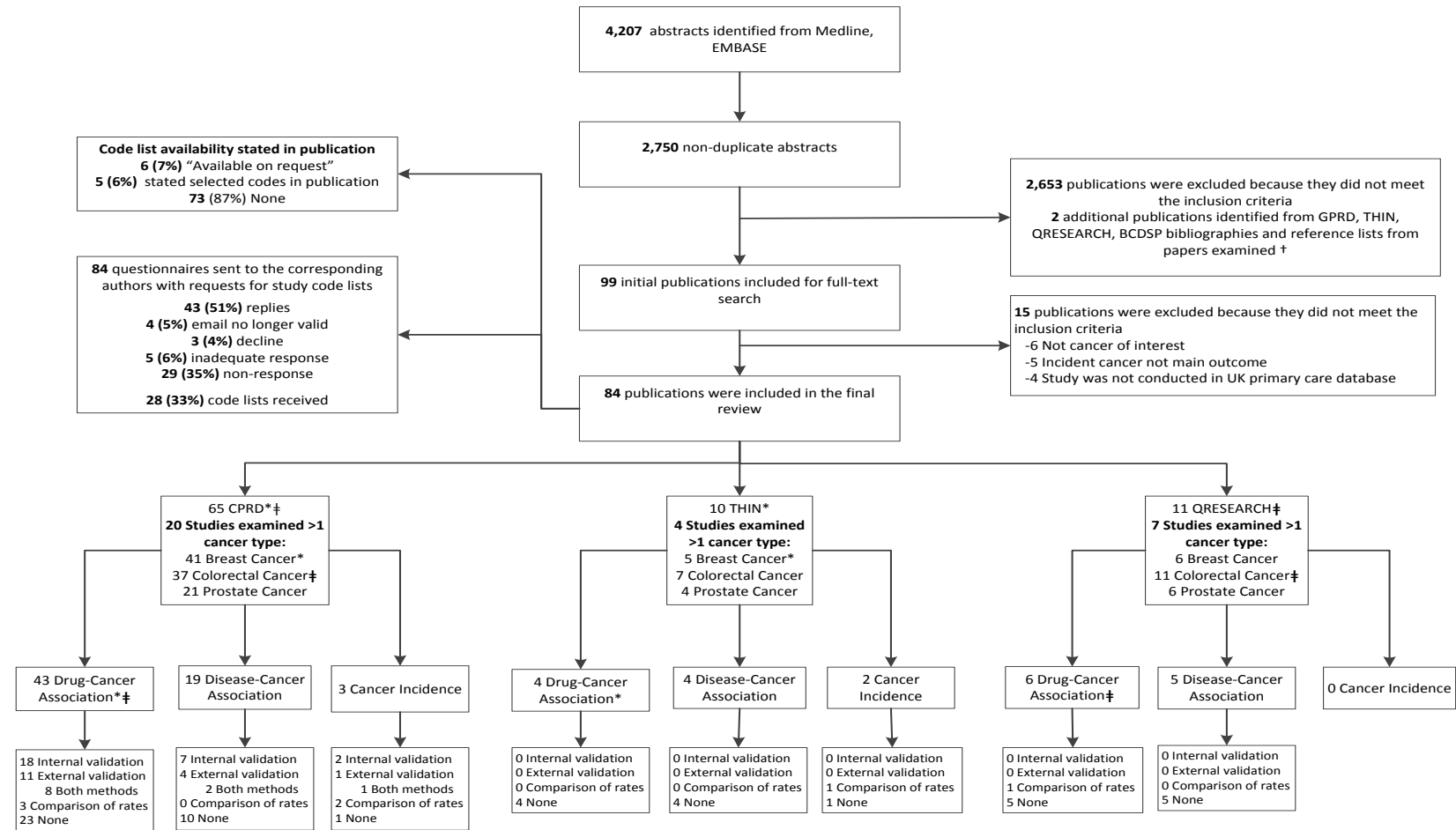
Overall, 84 relevant studies were included in this review (**Figure 2.1 and Appendix A, Table 8.1**). Studies utilised the CPRD (n=63); THIN (n=9); QRESEARCH (n=10); both the CPRD and THIN (n=1); and both the CPRD and QRESEARCH (n=1). Of the 84 studies, 30 examined >1 cancer types included in this review: breast (n=51); colorectal (n=54); and prostate cancer (n=31). A broad range of study types were included: 51 examined the association between drug use and cancer; 28 examined cancer incidence among patients with a particular disease or symptom; and 5 estimated population-level cancer incidence (**Figure 2.1**).

### **2.3.2 Study Code list Creation, Availability, and Comparison**

In total, only 5 of the 84 studies (6%) described methods used to create study code lists (**Table 2.1**). Five studies (6%) included details directly in the publication, 2 (2%) included the list itself as an appendix,<sup>80,81</sup> and 3 (4%) stated which Read code chapters and sections were used; a further 6 (7%) stated that the list was 'available on request'.<sup>63, 82, 83</sup>

Overall, there were 43 responses from 84 questionnaires sent to the authors (**Figure 2.1 & Table 2.1**). 37 (86%) studies reported using a keyword search of cancer related terms to identify potential cancer related codes, 26 (60%) utilised a previous code list, and 43 (100%) consulted with a health care professional during the creation of the study code list. For all studies, >1 assessor reviewed the code list.

**Figure 2.1: Flow diagram of article search, retrieval, and review process; code list availability and questionnaire replies; database, study type, and validation methods.**



† Boston Collaborative Drug Surveillance Program (BCDSP)

\* 1 study utilised both the Clinical Practice Research Datalink (CPRD) and The Health Improvement Network (THIN)

‡ 1 study utilised both the CPRD and QRESEARCH

In total, 28 of a potential 84 (33%) study code lists were received (**Figure 2.1**); frequencies of all codes included across the 28 studies are provided in **Appendix A, Table 8.2**. All 28 studies included malignant neoplasm diagnosis codes, but there was variation in the specific codes used: for breast cancer, 42 malignant disease codes were included across lists, but only 15 were included by all studies. The variation was not explained by changes in the Read code dictionary: all 42 codes were in the dictionary throughout the period when these studies were conducted (**Appendix A, Table 8.2**). Similar variation for colorectal cancer (64 malignant codes mentioned but only 18 appeared in all lists; all but 2 of the 64 codes were in the Read dictionary throughout), and for prostate cancer was found (8 malignant codes mentioned but only 1 appeared in all lists; all 8 codes present in the Read dictionary throughout).

There was also variability between lists in terms of other types of codes included: 20/28 (71%) code lists included in-situ tumours; 17 (61%) included malignant morphology codes; 20 (71%) included secondary or history of cancer codes; 16 (57%) included “borderline” codes; and 3 (10%) included suspected codes. In addition, a few lists included benign (n=5, 18%) and non-cancerous codes (n=4, 14%). It was not clear from the available information precisely how these various classes of codes were used for case ascertainment (**Table 2.1**).

**Table 2.1: Code list availability, questionnaire replies, and comparison of lists received, by cancer and study type**

	Cancer Type*				Study Type		
	All studies	Breast	Colorectal	Prostate	Drug-Cancer	Disease-Cancer	Incidence
	n (column %, n/N)						
<b>Total number of studies</b>	<b>N=84</b>	<b>N=51</b>	<b>N=54</b>	<b>N=31</b>	<b>N=51</b>	<b>N=28</b>	<b>N=5</b>
Any code list creation methods reported**	5 (6)	4 (8)	3 (2)	2 (6)	3 (6)	2 (7)	0 (0)
<b>Code list availability in publication</b>							
Available on request	6 (7)	2 (4)	6 (11)	2 (6)	3 (6)	2 (7)	1 (20)
Stated in publication	5 (6)	4 (8)	4 (7)	3 (10)	4 (8)	1 (4)	0 (0)
None	73 (87)	45 (88)	44 (81)	26 (86)	44 (86)	25 (89)	4 (80)
<b>Questionnaire results: number of replies</b>							
Keyword-synonym search	37 (86)	21 (88)	24 (80)	12 (80)	25 (83)	9 (90)	3 (100)
Utilisation of previous study code list	26 (60)	15 (63)	14 (47)	9 (60)	18 (60)	7 (70)	1 (33)
Consultation with health professional	43 (100)	24(100)	30 (100)	15 (100)	30 (100)	10 (100)	3 (100)
<b>Number of study code lists obtained</b>							
<b>Studies including specific code-types</b>	<b>N=28</b>	<b>N=17</b>	<b>N=23</b>	<b>N=11</b>	<b>N=21</b>	<b>N=5</b>	<b>N=2</b>
Malignant neoplasm	28 (100)	17 (100)	23 (100)	11 (100)	21 (100)	5 (100)	2 (100)
In-situ	20 (71)	12 (70)	15 (65)	6 (55)	13 (62)	5 (100)	2 (100)
Malignant morphology	17 (61)	13 (76)	11 (48)	- (-) <sup>†</sup>	12 (57)	3 (60)	2 (100)
Secondary or history of cancer	20 (71)	13 (76)	17 (74)	8 (73)	16 (76)	2 (40)	2 (100)
<b>Non-malignant codes</b>	<b>16 (57)</b>	<b>8 (47)</b>	<b>11 (48)</b>	<b>2 (18)</b>	<b>10 (48)</b>	<b>4 (80)</b>	<b>2 (100)</b>
Borderline codes	16 (57)	7 (41)	11 (48)	2 (18)	10 (48)	4 (80)	2 (100)
Suspected	3 (10)	2 (11)	2 (9)	1 (9)	1 (5)	1 (20)	1 (50)
Benign tumour codes	5 (18)	2 (11)	4 (17)	0 (0)	2 (10)	1 (20)	2 (100)
Non-cancerous or site-unrelated	4 (14)	1 (6)	4 (17)	0 (0)	1 (5)	2 (40)	1 (50)

\*One study could contribute to >1 cancer type

\*\*Code list creation methods include: keyword search of dictionary; review of code list by health professional; utilisation of previous code list

<sup>†</sup>There are no malignant morphology codes for prostate cancer found among the 11 prostate cancer studies

Stratification by study type indicated a possible difference in code inclusion between study types (**Table 2.1**). Both incidence studies included non-malignant codes (borderline, suspected, benign, and non-cancerous). Although lists were not received for 2 other incidence studies, they both stated using non-malignant codes within their publication.<sup>84,85</sup> In contrast, only 14/26 (56%) drug safety and epidemiological studies included non-malignant codes.

### **2.3.3 Identification and Validation of Cancers**

Of the 84 studies, the majority (n=57) only required  $\geq 1$  cancer diagnosis Read code to identify cases (**Table 2.2**). 27 studies specified additional criteria to confirm case status, for example, chemo-radiotherapy (n=13); biological treatment (n=12), surgical procedures (n=13). The requirement for further evidence was more common for breast cancer studies (76%) compared to colorectal (44%) and prostate cancer studies (50%). 11 studies mentioned a manual review process but did not report the criteria used to confirm or refute case status (**Table 2.2**). Where present, descriptions of diagnostic algorithms were typically brief; only one study provided a schematic of the algorithm used to identify and confirm case status.<sup>84</sup> Few studies (4/27) reported on the proportion of cases included once additional confirmatory evidence was applied. Gonzalez-Perez *et al.*<sup>86</sup> reported that 3708/3886 (95.4%) incident breast cancer cases had supporting evidence of diagnosis. Charlton *et al.*<sup>84</sup> identified 1,809 potential colorectal cancer cases, of which 1,599 patients (88.3%) had additional supporting evidence of diagnosis: colorectal cancer related surgery confirmed 927 cases (51.2%) and non-surgical support such as chemo-radiotherapy or palliative care confirmed 278 cases (15.4%). Of note, Bodmer *et al.*<sup>87</sup> assessed the effect of metformin on colorectal cancer incidence within the CPRD. Similar

estimates were obtained regardless of the requirement for confirmatory evidence of diagnosis (OR for  $\geq 50$  prescriptions vs. never use =1.43; 95% CI, 1.08-1.90 when cases were defined by codes alone and OR=1.46; 95% CI, 1.03-2.06 when restricting to those with further supportive evidence of cancer).

14 CPRD studies validated a sample of potential cases using information external to the database, namely by GP questionnaire or through a request of patient records (**Table 2.3**). The proportion of confirmed cases was high [median positive predictive value (PPV) = 0.99; Range, 0.90-1.00], although validity measures were limited to PPV. The number of potential cases sampled was low [median % 4.0; range, 0.8-11.1]. The median proportion of responses received was high [median proportion 0.95; range, 0.87-1.00]. External validation results stratified by cancer type were generally similar.



**Table 2.2: Criteria used to identify, validate, and exclude potential cancer cases by cancer and study type**

	Cancer Type*				Study Type		
	All studies	Breast	Colorectal	Prostate	Drug-Cancer	Disease-Cancer	Incidence
	n (column %, n/N)				n (column %, n/N)		
<b>Total number of studies</b>	<b>N=84</b>	<b>N=51</b>	<b>N=54</b>	<b>N=31</b>	<b>N=51</b>	<b>N=28</b>	<b>N=5</b>
<b>Number of studies requiring ≥1 cancer diagnosis code only</b>	57 (68)	34 (67)	45 (83)	23 (74)	33 (65)	21 (75)	3 (60)
<b>Internal validation or requirement for supportive evidence of diagnosis: number of studies</b>	<b>N=27</b>	<b>N=17</b>	<b>N=9</b>	<b>N=8</b>	<b>N=18</b>	<b>N=7</b>	<b>N=2</b>
Cancer related surgery	13 (48)	11 (65)	4(44)	3 (38)	8 (44)	3 (43)	2 (100)
Chemo/radiotherapy	13 (48)	10 (59)	4 (44)	4 (50)	7 (39)	4 (57)	2 (100)
Biological treatment	12 (44)	10 (59)	2 (22)	4 (50)	8 (44)	3 (43)	1 (50)
Treatment unspecified	1 (4)	1 (6)	0 (0)	0 (0)	1 (6)	0 (0)	0 (0)
Consultation with oncologist	8 (30)	7 (41)	2 (22)	1 (13)	6 (31)	1 (14)	1 (50)
Other <sup>†</sup>	3 (11)	2 (12)	3 (33)	1 (13)	1 (6)	1 (14)	1 (50)
Unspecified <sup>‡</sup>	11 (41)	4 (24)	5 (56)	4 (50)	8 (44)	3 (43)	0 (0)
<b>Cancer related exclusion criteria: number of studies</b>							
Previous diagnosis of any cancer	43 (51)	31 (61)	25 (46)	19 (61)	29 (57)	12 (43)	2 (40)
Previous diagnosis of cancer of interest	25 (30)	10 (20)	18 (33)	6 (19)	16 (31)	6 (21)	3 (60)
Time related exclusion periods	59 (70)	32 (63)	39 (72)	24 (77)	37 (37)	19 (68)	3 (60)

\*One study could contribute to >1 study type

<sup>†</sup>Other includes: specific oncology codes, terminal illness, palliative care, death within 180 days of diagnosis.

<sup>‡</sup>A manual review process was conducted, however criteria used to confirm case status was not described.

**Table 2.3: External validation of potential cases by cancer type**

	Cancer Type*			
	All studies	Breast	Colorectal	Prostate
	n (column %), median [Range], unless otherwise specified			
<b>Total number of studies</b>	<b>N=84</b>	<b>N=51</b>	<b>N=54</b>	<b>N=31</b>
<b>Number of studies that validated cases externally by questionnaire or request for patient records (%)<sup>‡</sup></b>	14 (20)	5 (12)	4 (10)	3 (14)
Number of potential cases sampled for external validation	100 [23-200]	114 [30-114]	85 [23-200]	100 [100-100]
Proportion of cases randomly sampled for external validation from patients initially fulfilling inclusion criteria	4.03 [0.81-11.06]	3.07 [0.81-3.07]	6.40 [3.49-11.06]	7.21 [4.58-9.85]
Proportion of responses received	0.95 [0.87-1.00]	0.95 [0.95-1.00]	0.96 [0.87-1.00]	- <sup>‡</sup>
Proportion of cases confirmed	0.99 [0.90-1.00]	1.00 [1.00-1.00]	0.95 [0.90-1.00]	0.98 [0.98-0.98]
<b>Number of studies that validated cases externally by linkage to cancer registry (%)</b>	<b>N=2</b>	<b>N=1</b>	<b>N=2</b>	<b>N=1</b>
Number of potential CPRD cases sampled for external validation	703 [-]	560 [-]	1228 [681-1775]	725 [-]
Proportion of cases confirmed in cancer registry median [range]	0.90 [0.83-0.94]	0.90 [-]	0.94 [0.91-0.98]	0.83 [-]

\*One study could contribute to >1 cancer type

<sup>‡</sup>Only one study reported the number of responses received – Ronquist *et al*: 88 responses received from a request of 100 patient records

<sup>‡</sup>Two studies included in “All studies”, but were not included in specific cancer type columns as they externally validated overall cancer - not distinguishing by cancer type

Two studies examined the concordance of recorded cancer diagnosis between the CPRD and UK cancer registry.<sup>80, 88</sup> Estimates of concordance between the CPRD and cancer registry were high [median PPV 0.9; range, [0.8-0.9]. Dregan *et al.*<sup>80</sup> reported a PPV of 0.98 for colorectal cancer and Boggon *et al.*<sup>88</sup> reported similarly high PPVs for cancer of the breast (503/560=0.90), prostate (600/725=0.83); and colorectum (618/681=0.91). Sensitivity estimates of the CPRD in capturing registered cancers estimates were also high: Boggon *et al.*<sup>88</sup> reported sensitivity estimates ranging from 95% for colorectal cancer to 99% for prostate cancer; similarly, Dregan *et al.*<sup>80</sup> reported 92% sensitivity for colorectal cancer in capturing cancer registry recorded diagnoses.

#### **2.3.3.1 Comparison of cancer incidence rates**

A database-level method of validation can be applied by comparing database cancer incidence rates to incidence rates from a reputable external source. Seven studies compared database specific incidence rates of cancer diagnosis to an external data source, of which three reported lower database incidence rates compared to published rates estimated by the Office for National Statistics (ONS) (**Table 2.4**). Seven studies compared cancer incidence rates to different external data sources. Four studies found similar incidence rates (2 breast cancer, 1 colorectal, 1 prostate), while three studies reported lower colorectal cancer incidence rates when compared to external data sources. Colorectal cancer incidence rates were reported by four studies with conflicting findings.<sup>84, 89-91</sup> A recent study by Charlton *et al.*<sup>84</sup> reported lower colorectal cancer incidence rates in the CPRD (incidence rate per 100,000 person years (100k PY); Men: 63.7, Women: 48.4) compared to UK cancer registries (year 2007: men, 70.2; women, 56.6 per

100,000 PY). Similarly, Vinogradova *et al.*<sup>90</sup> reported an overall incidence rate of 49.8 per 100,000 PY (men: 56.1, women: 43.6 per 100,000 PY) in the QRESEARCH database which was lower compared to published ONS incidence rates in 2003 (men: 62.3; women 49.5 per 100,000 PY). In contrast, Garcia Rodriguez *et al.*<sup>91</sup> reported an overall CPRD incidence rate of 73 per 100,000 PY for colorectal cancer. However, despite the study period being between 1994-1997, the overall rate was significantly higher compared to the estimated rates observed by Charlton *et al.*<sup>84</sup> and Vinogradova *et al.*<sup>90</sup> and the National Cancer Intelligence Network between 1995-2004 (men: 62.3; women: 53.4).<sup>92</sup> Of note, 6 of the 7 studies compared crude cancer incidence rates (both estimated from the UK primary care database and crude estimates reported by the ONS). Only 1 study compared age-standardised incidence rates estimated from THIN to equivalent age-standardised rates reported by the ONS.

**Table 2.4: Comparison of incidence rates by cancer type**

	Cancer Type*			
	All studies	Breast	Colorectal	Prostate
	n (column %), median [IQR], or otherwise specified			
<b>Total number of studies</b>	<b>N=69</b>	<b>N=41</b>	<b>N=42</b>	<b>N=22</b>
<b>Comparison of Incidence Rates (IR)**</b>				
No. of studies comparing database IR to an external data source	7 (10)	2 (5)	4 (10)	1 (5)
IR per 100,000 person years [range]	-	156.0 [-]	49.8 [49.5-73.0]	161 [-]
Result of IR comparison: higher, lower, similar †	4 similar; 3 lower	2 similar <sup>85, 93</sup>	1 similar; <sup>94</sup> 3 lower <sup>84, 90, 95</sup>	1 similar <sup>96</sup>

\* One study can contribute to >1 cancer type

† Incidence rate percentage differences could not be estimated due to non-reporting of database incidence rates by two studies investigating colorectal<sup>95</sup> and breast cancer<sup>85</sup>

\*\* 6 of the 7 studies reported crude cancer incidence rates. Only Haynes *et al*<sup>95</sup>. reported age-standardised incidence rates

## **2.4 Discussion**

### **2.4.1 Overview**

This review has revealed several common shortcomings related to the description of methods used to identify cancer cases in UK primary care database studies. Few studies reported the methods used to compile code lists, or made code lists available, limiting the reproducibility of studies. Furthermore, where information was available, substantial variation in codes included was observed. High positive predictive value estimates were reported for all three cancer types from studies that used information external to the database to validate cases, but other measures of validity such as sensitivity and specificity were not generally explored.

### **2.4.2 Accessibility of code lists**

Only 11/84 studies made their code lists available in the publication or specifically mentioned that they could be requested. Code lists may not have been made available for several reasons. For the earlier studies included, there may have been no practical way of publishing a long code list. More recently, most journals have started accepting web appendix materials without space limits, and other alternatives have emerged, such as including a web link in the paper to a central code lists repository or registry of studies. Making code lists 'available on request' is problematic since there may be difficulties in contacting the original corresponding author, particularly as time elapses after publication. Some authors simply may not have considered code lists to be important supplementary information, suggesting a need to raise awareness of the need for clear reporting of case definitions. Lastly, there may be some reluctance among researchers to release code lists due to

concerns that they could be used by competing research groups and without due credit.

### **2.4.3 Variation in case definitions and code lists**

There was considerable variation in the specific codes used by researchers to identify cancers. The Read code dictionary is updated regularly but was not found to be an important driver of variation between code lists; the vast majority of codes used by investigators were available throughout the period during which the included studies took place. It is worth noting that variation in code lists does not necessarily translate to an equivalent variation in selected cases, which will also depend on how commonly specific codes are used. For example, if a majority of cases of breast cancer have a Read code for 'Malignant Neoplasm of Female Breast' (B34..00, which was included in all code lists for reviewed breast cancer studies) then these cases will be identified regardless of content in the rest of the code list. In the other direction, including a code which is never used in practice will have no effect on case ascertainment.

As well as variation in individual codes, variation in types of codes included was observed. All lists included definite malignant diagnosis codes, but some included other code types such as carcinoma in-situ or suspected cancer. Some of the variation in definitions is likely to have arose from differing study objectives; differences by study type was noted, as may be expected. For example, pharmacoepidemiological studies aiming for high specificity may only include definite malignant neoplasms and exclude borderline codes.<sup>63, 97</sup> While incidence studies may use a broad code list to maximise sensitivity, and then attempt to confirm diagnosis in a second stage of review.<sup>84, 85</sup> Some studies included benign

tumour and non-cancer codes without explanation; whether such codes were included mistakenly or were used specifically to exclude cases is unclear.

The majority of studies required only a cancer diagnosis code as part of their case definition, but around a third of studies required some form of further supportive evidence to confirm case status. Again, there were limited details in many study reports on the specific diagnostic algorithms used; one study presented a full schematic illustrating the case definition algorithm,<sup>84</sup> routine use of such diagrams might help to improve clarity.

#### **2.4.4 External validation of cancer cases**

A number of studies validated cases externally by request of patient records or by GP questionnaire and were generally able to confirm a high proportion of cases. However, not all practices participate in validation or linkage studies, which may limit the generalisability of validity findings if participating practices differ from non-participating practices in terms of record-keeping practices. It is also unclear whether GP practices asked to validate cases in this way are accessing extra information, or simply referring to the same electronic record, which would inevitably lead to optimistic validity estimates.

#### **2.4.5 Comparison of incidence rates**

Few studies compared database specific cancer incidence rates to ONS published incidence rates. Disparities by age and calendar year were observed for colorectal cancer, while similar rates were observed for breast and prostate cancer. Although disparities were observed, the majority of compared incidence rates reported (6 out of 7 studies) were crude and not age-standardised which may limit comparative interpretations if age distributions differ between data sources. That being said, age



and sex distributions have been shown to be representative of the UK population.<sup>98</sup>

Another possible validation method is the comparison of survival/mortality estimates from UK primary care databases to national estimates. Only Boggon *et al.*<sup>88</sup> compared survival estimates in the CPRD to that of the cancer registry among a cohort of diabetic patients. Boggon *et al.*<sup>88</sup> observed consistently higher survival estimates in primary care compared to linked cancer registry estimates. Breast cancer had the largest difference among 11 other cancer types, suggesting discrepancies in terms of diagnosis dates between the two data sources.

Importantly, comparison of cancer survival estimates would capture problems with the recording of incident cancer diagnosis, date of diagnosis, cause of death, and date of death recording. However, it may be difficult to identify the exact cause(s) of the discrepancy based on a survival estimate alone.

#### **2.4.6 Limitations of this review**

This review had several limitations: firstly, results were limited to cancers of the breast, colorectum, and prostate, and may not apply to other malignancies.

Nonetheless, many of the studies included in this review examined multiple cancers, and applied case ascertainment in a global fashion rather than separately for each cancer type. Secondly, this review was limited to UK primary care database studies; whether the variation observed in this review occurs in non-UK databases is unknown. Lastly, authors who completed questionnaires and sent code lists may have been a selective group; therefore, their responses may not be generalisable to all researchers. Non-response or an unwillingness to share code lists may have arisen due to concern about methodological criticism, or protectiveness over intellectual property.

### **2.4.7 Importance and implications**

Our study highlights the variation and lack of transparency in many studies to date on a critical methodological feature of database studies of cancer outcomes, namely the definition and ascertainment of cancer cases. Primary care databases and routine healthcare records are increasingly used in cancer research. A total of 84 relevant articles covering just 3 cancer sites; a broader search not restricted by site finds >250 articles published in leading general medical journals and influential specialist journals. Clarity over case ascertainment methods is important for interpreting study findings, reproducing analyses, and understanding the drivers of discrepant results.<sup>19, 99</sup> Recent work by the Observational Medical Outcomes Partnership has highlighted that design decisions in observational pharmacoepidemiology studies profoundly affect study results,<sup>100</sup> further emphasising the importance of clear and transparent reporting. As well as directly highlighting the need for such transparency and thus influencing future studies, this work can also inform guidelines aimed at improving the quality of reporting for electronic healthcare record research, which are currently in development as part of the RECORD project.<sup>101</sup>

### **2.4.8 Updated review studies**

An update of the systematic review was conducted to examine the current findings from studies undertaken after the original systematic review. Twelve studies were identified from a re-run of the original literature search in July 2015. Six studies examining breast cancer,<sup>102-107</sup> 6 studies colorectal cancer,<sup>102, 103, 105, 107-112</sup> and 6 studies prostate cancer (a study could examine >1 cancer type).<sup>102, 103, 105-107, 113</sup> Of the 13 studies, 2 made their code list available through a published appendix,<sup>102, 103</sup> and one study mentioned excluding chapter B7 from the final code list. Most studies (N=11) defined cancer by requiring one malignant diagnosis code and only one study implemented a diagnostic algorithm to identify cases of colorectal cancer. Findings from the additional studies were generally consistent with the original systematic review.

## **2.5 Conclusion**

This review comprehensively investigated several aspects of case ascertainment from studies utilising primary care databases for research related to cancer. Methods used to develop case definitions were often unclear and specific code lists were seldom published or made available. Where provided, considerable variation in case definitions and code lists was observed, and its impact on case ascertainment is unclear. Future research might clarify the extent to which methodological variations identified in this review impact on findings in applied epidemiological studies, and further explore ways of validating cancer case definitions, including through the use of linked data sources and free-text information.<sup>80, 88, 114</sup> It is hoped that this study will help to promote clearer reporting of cancer case ascertainment methods, better access to code lists, and a

resulting improvement in the transparency and reproducibility of research in this growing field.

## **2.6 Addition of lung cancer to cancer outcomes of interest: rationale**

This review concentrated on breast, colorectal, and prostate cancer. However, from this review there is evidence suggesting fatality may influence case identification in primary care. In comparison to national rates reported by the ONS, primary care incidence rates for colorectal cancer were lower (colorectal cancer 1-year survival, 76%). Although data on breast (1-year survival, 96%) and prostate cancer (1-year survival, 94%) were limited, similar rates were observed from primary care compared to ONS reported rates for the two cancer types. In order to assess the impact of fatality on case identification and conflicting findings between studies, lung cancer (1-year survival, 32%) was added to the cancer outcomes initially planned for investigation in this thesis.

## **2.7 Summary**

- Methods used to create outcome code lists were not transparent in the majority of studies included in this review, and the overall accessibility of study code lists was low.
- Substantial variation in the way cancer cases were defined was observed, including in the specific diagnosis codes used, and the requirements for further confirmatory evidence. This could potentially impact case ascertainment and study findings.
- Cancer outcomes defined using database-recorded information had high positive predictive value, when validated against external data sources, but

few data were available on other measures of validity such as sensitivity and specificity.

- Compared to national estimates, lower primary care incidence rates were observed in three studies. However, reasons for the disparity are unclear.
- Transparency and reproducibility of research would be improved by clearer reporting of methods used to develop case definitions, and by making code lists available for all published studies.

### **3 Systematic review of conflicting findings in observational studies utilising electronic patient records to investigate statin use and cancer risk**

#### **3.1 Introduction**

This chapter reviews existing observational studies that have utilised electronic healthcare records to investigate the association between statin use and the risk of breast, colorectal, lung, and prostate cancer.

#### **3.2 Aims**

The aims of this chapter were to collate published literature and to:

1. Describe and summarise the results of observational studies utilising electronic healthcare records to investigate the association between statin use and the risk of breast, colorectal, lung, and prostate cancer.
2. Investigate the discrepancies between studies, and identify potential methodological limitations that might contribute to inconsistent findings.

#### **3.3 Methods**

##### **3.3.1 Databases and sources**

Medline and EMBASE were searched for abstracts published between January 1, 1987–November 30, 2011 using key words and related synonyms. In addition, reference lists of review studies were searched for articles that may have been missed by the database searches. Conference abstracts and unpublished studies were excluded from this review.

##### **3.3.2 Search keywords and terms**

The search of MEDLINE (via OvidSP) included exploded key terms to identify publications that investigated the association between statin use and specific cancer

types: breast, colorectum, lung, and prostate. For EMBASE, which does not use the MeSH classification system, the nearest equivalent search terms from the EMBASE indexing system was used.

The following MeSH keywords were used in the primary search:

[\*Statin OR Hydroxymethylglutaryl-CoA Reductase OR \*CoA OR Simvastatin (plus all sub-terms in the MeSH tree)]

**AND**

[Cancer OR Neoplasm OR Malig\*(plus all sub-terms in the MeSH tree)]

**AND**

[Epidemiologic OR observational OR Case-control OR Cohort OR Retrospective (plus all sub-terms in the MeSH tree)];

**AND**

Limited to article types: JOURNAL ARTICLE; limited to subjects: HUMANS; limited to language: ENGLISH

### **3.3.3 Inclusion and exclusion criteria**

#### **Inclusion criteria:**

1. Observational study.
2. Utilisation of routinely collected electronic patient data.
3. Incident cancer of the breast, colorectum, lung or prostate as the primary outcome.
4. Statin drug use as the main exposure of interest.
5. Study on humans.
6. Manuscript in English.

#### **Exclusion criteria:**

1. Conference abstracts.
2. Review articles.

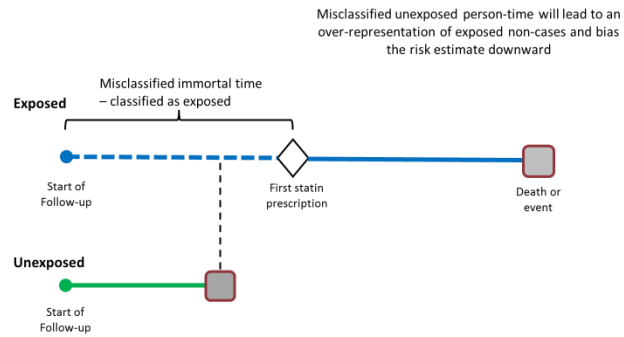
**Table 3.1: Description of biases assessed**

Bias	Description of the bias	Direction of bias	Methods to minimize bias
<b>Immortal time bias</b>	Immortal time bias refers to a period of follow-up time during which, by design, death or the study outcome cannot occur. <sup>69</sup> This bias arises when the treatment status of a patient is defined by a set criterion of minimum exposure or a wait period during which follow-up time is accrued.  Bias is introduced when this unexposed wait period e.g. cohort entry until first prescription (immortal time – since a patient has to survive to first prescription in order to be classified as exposed) is misclassified as exposed. As exposed subjects include unexposed immortal time, risk estimates may be biased downward in favour of the exposed group.	Downward	<ul style="list-style-type: none"> <li>• Analysis incorporating a time dependent exposure (Poisson or Cox proportional hazards regression).<sup>115</sup></li> <li>• Exclusion of immortal time bias.<sup>115</sup></li> </ul>
<b>Protopathic bias</b>	Protopathic bias can occur if symptoms of a pre-existing cancer are associated with patients being prescribed statins by their GP, which can lead to an artificial increase in cancer risk. <sup>71</sup>  In the other direction, patients who are ineligible to receive statins, or discontinue statin therapy due to cancer related symptoms, may lead to a falsely low rate of statin usage in patients who have cancer, which could bias risk estimates downward. <sup>116</sup>	Upward or downward	<ul style="list-style-type: none"> <li>• Lag-time prior to index date to assure minimum period of exposure for case-control studies.<sup>116</sup></li> <li>• Minimum period of exposure for cohort studies.</li> </ul>
<b>Prevalent user bias</b>	Prevalent users of statins may differ compared to incident statin use ( <i>new</i> users). Differences between the two groups of patients may occur due to various factors such as adherence and tolerance to medications; attendance to health utilisation services and cancer screening/prevention programs, which may bias risk downward if prevalent users are included in study design.	Upward or downward	<ul style="list-style-type: none"> <li>• <i>New user design</i>.<sup>70</sup></li> </ul>
<b>Healthy user bias</b>	Users of preventative therapy such as statins or antihypertensive medications may exercise more, have a healthier diet, and may adhere to health services directed at preventing related diseases compared to the general population. <sup>117</sup>  Healthy user bias may increase the likelihood of cancer detection among statin users, and can increase the incidence rate of particular cancers, and hence bias risk estimates upward among statin users compared to non-statin users from the general population.  Healthy user bias may bias results downward if confounders such as diet, physical activity, or adherence are not adjusted within statistical analysis.	Upward or downward	<ul style="list-style-type: none"> <li>• Comparison group consisting of patients prescribed a preventative therapy. This group of patients may be similar to statin users in terms of adherence to medications and healthy behaviour.<sup>118</sup></li> <li>• Adjustment for lifestyle factors and access to health utilisation services.<sup>118</sup></li> </ul>
<b>Time-window bias</b>	Time-window bias arises when the time-period used to assess exposure status between cases and controls is not equal, and the sampling of controls is biased. <sup>119</sup> If sampled at their last point of contact, controls may have a longer follow-up period and hence a higher likelihood of exposure being observed compared to cases, resulting in a spurious appearance of a benefit of exposure.	Downward	<ul style="list-style-type: none"> <li>• Matching on duration of follow-up.<sup>119</sup></li> <li>• Risk-set sampling.<sup>119</sup></li> </ul>

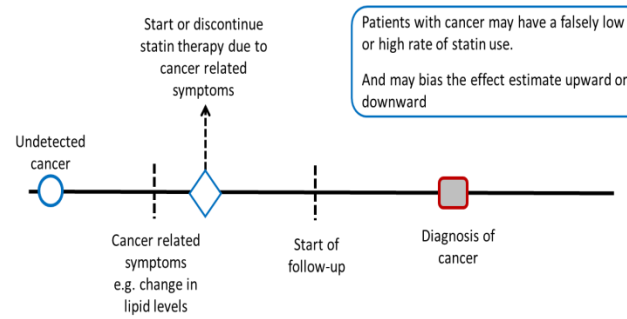


**Figure 3.1: Diagrammatic illustration of biases examined**

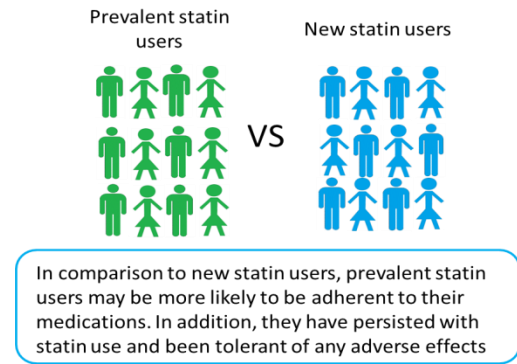
**(a) Diagram of immortal time bias**



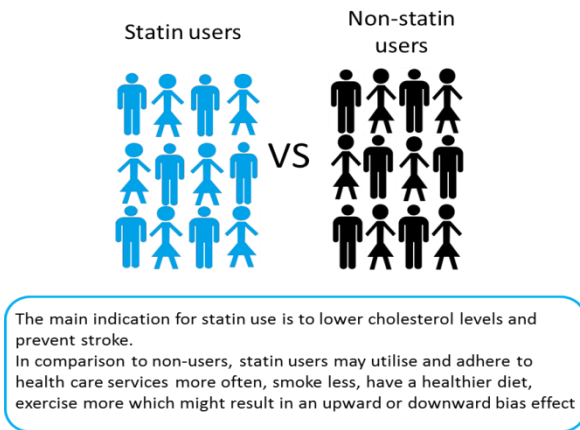
**(b) Diagram of protopathic bias**



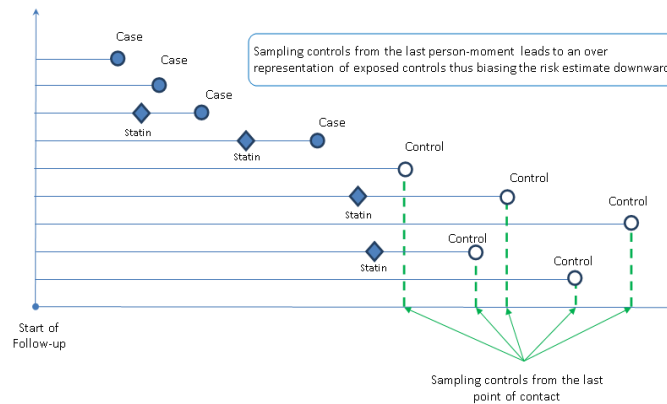
**(c) Diagram of prevalent user bias**



**(d) Diagram of healthy user bias**



**(e) Diagram of time-window bias**



### 3.3.4 Procedure

All publications were reviewed; extracted data included: source population, data sources, primary and secondary analysis results. Methodological features were also recorded and included: case and exposure definitions; comparison groups; control selection; adjustment for confounders; and statistical methods used.

### 3.3.5 Bias selection

Five biases were assessed, including: immortal time, protopathic, prevalent user, healthy user, and time-window bias. Selection of these biases were based on discussion within the advisory group about key suspected biases in pharmacoepidemiology, as well as reviewing the literature on statin use and cancer risk which suggested that these particular biases may be important. Although, not all biases were shown to systematically impact across studies of different drug-disease associations<sup>69, 70, 116, 117, 119-122, 123, 124</sup>

Each bias is described in **Table 3.1** and illustrated in **Figure 3.1**. Each paper was assessed for the possibility of these biases, as follows:

**Immortal time bias** was considered avoided if (i) a time-dependent analysis of statin treatment was implemented; or if (ii) immortal time periods were excluded from both exposure and referent groups. Assessment of this bias was restricted to cohort studies.

**Protopathic bias** was considered mitigated if minimum periods of exposure or lag time periods were implemented within primary, secondary or sensitivity analyses.

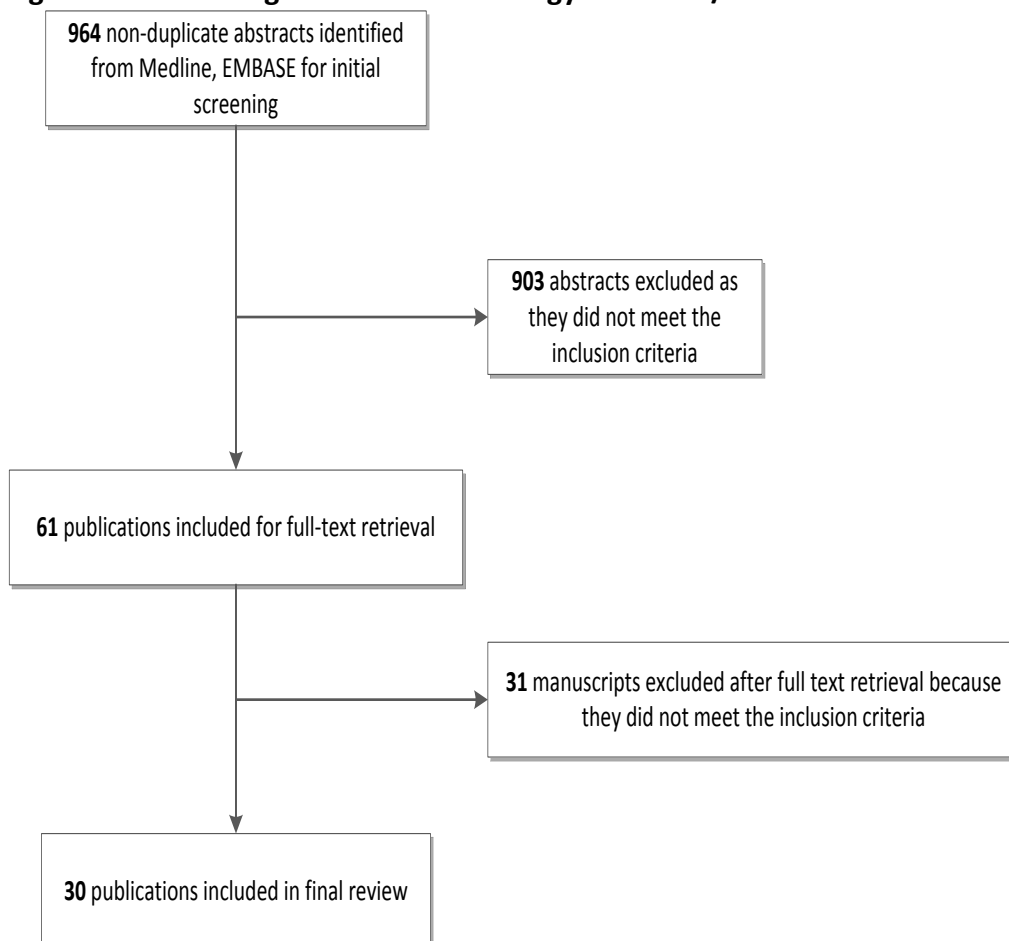
**Prevalent user bias** was considered minimised if a *new* user design was implemented.

**Healthy user bias** was considered minimised if (i) a comparison group consisting of patients prescribed preventative therapy medications was used; or (ii) there was an

adjustment for uptake of preventative health services such as GP/physician consultations, hospitalisations; cancer related examinations such as screening, prostate specific antigen (PSA) testing, colonoscopy, stool occult blood test, digital rectal examination, mammogram.

**Time-window bias** was minimised if the study authors implemented a case-control design in which controls were selected by risk-set sampling or if they were matched on follow-up or date of diagnosis. Assessment of this bias was restricted to case-control studies.

**Figure 3.2: Flow diagram of search strategy: inclusion/exclusion criteria.**



### 3.3.6 Meta-analysis

Firstly, a meta-analysis to compare the risk of breast, colorectal, lung and prostate cancer risk among statin users compared to non-users was conducted by using a

DerSimonian-Laird random effects model,<sup>125</sup> accounting for heterogeneity among studies, was used to calculate summary relative risk (SRR) and 95% CIs. The possible heterogeneity among studies was examined by using the  $I^2$ , the percentage of total variation across studies that is due to heterogeneity.<sup>126</sup>

Sources of heterogeneity were assessed by subgroup analyses according to study design (cohort vs case-control); comparison group (non-user vs other); new user vs any user; adjustment for smoking status; geographical location (European and Asian studies). If no more than 3 studies reported the association between a specific cancer type and exposure/referent, the overall effect was not summarised. For studies that only reported results for men and women, statin type separately, they were considered as independent data obtained from different studies.

### **3.4 Results**

An overview of overall results is first presented; methodological considerations are then described. Second, results by cancer type are described, and then an assessment of bias in the context of these results are given.

#### **3.4.1 Overview**

A total of 30 relevant studies were identified for this review (**Figure 3.2**). Of the 30 observational studies investigating the effect of statin use on breast, colorectal, lung or prostate cancer, 11 examined >1 cancer type: breast (n=14), colorectal (n=17), lung (n=11), and prostate (n=17) (**Table 3.2 and Figure 3.3**). 16 studies considered a cohort design and 14 a case-control design. There were variations in design methodology applied particularly in terms of treatment definitions, and adjustment for potential confounders.

### 3.4.2 Methodological considerations

Variations in methodological aspects between studies occurred in several areas:

populations, outcome ascertainment, outcome definition, exposure definition, comparison groups, and confounder adjustment (**Table 3.3 and Table 3.4**).

#### 3.4.2.1 Populations

Studies included in this review were conducted in a broad range of populations and electronic health data sources (**Table 3.3**). Thirteen studies utilised administrative

healthcare data from various healthcare programs in the US, including: Kaiser

Permanente (KP), California (n=4); Veterans affairs (VA) healthcare system (n=4);

Group Health, Washington healthcare system (n=3); and resident patients ≥65 years

from Pennsylvania state (n=1); resident patients attending Cleveland Clinic (n=1).

Seven studies utilised data from three UK primary databases: The CPRD,<sup>127-129</sup>

QRESEARCH,<sup>90, 130, 131</sup>; and THIN.<sup>56</sup> The remaining 10 studies were conducted in various

populations including Finnish population registry (n=3); Canada: Quebec (n=1), the

provinces of Saskatchewan insurance claim (n=1) and Manitoba (n=1); Danish registries

(n=2); The Netherlands (n=1); and the Taiwanese National Health Insurance

programme (TNIH; n=1).

**Table 3.2: Frequency (%) of findings by cancer type**

Cancer	No significant association	Any statin use		Total
		Reduced risk	Increased risk	
Prostate n (%)	10 (59)	4 (24)	3 (18)	17 (100)
Colorectal n (%)	13 (76)	3 (18)	1 (6)	17 (100)
Breast n (%)	14 (100)	0 (0)	0 (0)	14 (100)
Lung n (%)	6 (55)	3 (27)	2 (18)	11 (100)

**Table 3.3: Statin use associated with cancer risk - study details**

Author & Year	Cancer(s) examined	Study Population	Data source for outcomes	Invasive or non-invasive cancer	Data source for statin use	Time period	Study design	Comparison group	Statin exposure definition	No. of cases (cases exposed)	Total cancer or site-specific cancer: point estimate (95% CI)
Chang 2011 <sup>68</sup>	Prostate	Taiwan, National Health Insurance program	National Health Insurance database	Invasive	Prescription database	1996-2008	Case-control	Non-Statin users	≥1 statin prescription at any time during the study period	388 (83)	1.55 (1.09-2.19)
Farwell 2011 <sup>132</sup>	Prostate	Veterans Affairs New England Healthcare system	VA database	Invasive	Prescription database	1997-2007	Cohort	Antihypertensive drug users	≥1 statin prescription; 2-year from date of first statin	546 (359)	0.69 (0.52-0.90)
Tan 2011 <sup>133</sup>	Prostate	Patients who underwent a prostate biopsy at Cleveland clinic	Clinic electronic records	Uncertain	Prescription database	2000-2010	Case-control	Non-Statin users	≥1 statin prescription prior to follow-up	2407 (565)	0.92 (0.85-0.98)
Vinogradova 2011 <sup>130</sup>	Prostate Colorectal Breast Lung	UK general practices (QRESEARCH database)	GP Records	Invasive	GP Prescription records	1998-2008	Nested Case-control	Non-Statin users	≥2 prescriptions at any time during the study; 1-year lag from date of diagnosis	88125 (13621)	1.01 (0.99-1.04)
Hippisley-Cox 2010 <sup>131</sup>	Prostate Colon Breast Lung	UK general practices (QRESEARCH database)	GP Records	Uncertain	GP Prescription records	2002-2008	Cohort	Non-Statin users	New user: ≥1 prescription at any time during the study period	_*	No association*
Murtola 2010 <sup>134</sup>	Prostate	Screening arm of the Finnish prostate cancer screening trial	Finnish cancer registry	Yes	Prescription database	1996-2004	Cohort	Non-Statin users	≥1 statin prescription at any time during the study period	1594 (268)	0.75 (0.63-0.89)
Robertson 2010 <sup>135</sup>	Colorectal	Counties of Aarhus and North Jutland, Denmark National Health Service	Danish National Registry (hospital records)	Invasive	Prescription database	1991-2008	Case-control	Non-statin users	≥2 prescriptions at any time during the study	9979 (711)	0.87 (0.80-0.96)
Wooditschka 2010 <sup>136</sup>	Breast	Kaiser Permanente (KP) Northern California members	KP cancer registry	Invasive	Prescription database	1997-2007	Case-control	Non-Statin users	≥2 prescriptions and ≥2 years of statin use	22488 (509)	1.02 (0.97-1.08)
Flick 2009 <sup>137</sup>	Colorectal	KP Northern and Southern California members	KP cancer registry	Invasive	Prescription database	2002-2003	Cohort	Non-Statin users	>100 day supply of one or more statins	171 (56)	0.89 (0.61-1.30)
Hachem 2009 <sup>138</sup>	Colorectal	Diabetic Veterans (VA database with linkage to Medicare patient files)	VA and Medicare database	Invasive and non-invasive	Prescription database	1997-2002	Nested case-control	Non-Statin users	≥1 statin prescription at any time during the study period	6080 (2987)	0.88 (0.83-0.93)
Hauka 2009 <sup>67</sup>	Prostate Colorectal Breast Lung	Finish cancer registry and social insurance institution	Finnish cancer registry	Uncertain	Prescription database	1996-2005	Cohort	Non-Statin users	≥1 prescription at any time during the study period	50294 (25445)	1.00 (0.98-1.02)
Singh 2009 <sup>64</sup>	Colorectal	Manitoba Health and Healthy Living insurance provider	Manitoba cancer registry	Uncertain	Prescription database	1995-2005	Cohort	Non-Statin users	≥2 prescriptions at any time during the study with no gaps >90 days	6637 (402)	1.13 (1.02-1.25)

[continued over]

[Table 3.3 continued]											
Author & Year	Cancer(s) examined	Study Population	Data source for outcomes	Outcome definition	Data source for statin use	Time-period	Study design	Comparison group	Statin exposure definition	No. of cases (cases exposed)	Total cancer or site-specific cancer: point estimate (95% CI)
Boudreau 2008 <sup>139</sup>	Colorectal	Group health (health care system) Washington	Washington cancer registry	Invasive and non-invasive	Prescription database	2000-2003	Case-control	Non-Statin users	≥2 prescriptions within any 6-month period and statin use for >1 year	357 (60)	1.02 (0.65-1.59)
Boudreau 2008 <sup>140</sup>	Prostate	Group health (health care system) Washington	Washington cancer registry	Invasive	Prescription database	1990-2005	Cohort	Non-Statin users	≥2 prescriptions within any 6-month period and statin use for >1 year	2532 (246)	0.88 (0.76-1.02)
Farwell 2008 <sup>141</sup>	Prostate Colorectal Lung	VA New England Healthcare system	VA database	Invasive	Prescription database	1997-2005	Cohort	Antihypertensive drug users	≥1 statin prescription; 2-year from date of first statin	6896 (2515)	0.74 (0.70-0.78)
Friedman 2008 <sup>142</sup>	Prostate Colorectal Breast Lung	Kaiser Permanente Medical care program Northern California	KP cancer registry	Invasive	Prescription database	1994-2003	Cohort	Non-Statin users	≥1 prescription at any time during the study period	Women: 2694 (-) Men: 4195 (-)	1.03 (0.99-1.07)
Smeeth 2008 <sup>56</sup>	Prostate Breast	UK general practices (THIN database)	GP Records	Uncertain	GP Prescription records	1995-2006	Cohort	Non-Statin users	New user: First prescription on or after Jan 1995 and >12 months continuous registration with a general practice	26484 (2471)	1.03 (0.96-1.11)
Yang 2008 <sup>127</sup>	Colorectal	UK general practices (CPRD database)	GP Records	Uncertain	GP Prescription records	1987-2002	Nested case-control	Non-Statin users	≥5 years continuous statin use	4432 (-)	1.1 (0.5-2.2)
Boudreau 2007 <sup>143</sup>	Breast	Group health (health care system) Washington	Washington cancer registry	Invasive	Prescription database	1990-2004	Cohort	Non-Statin users	≥2 prescriptions within any 6-month period and statin use for >1 year	861 (-)	0.90 (0.7-1.2)
Flick 2007 <sup>144</sup>	Prostate	KP Northern and Southern California	KP cancer registry	Invasive	Prescription database	California, 2002-2003	Cohort	Non-Statin users	>100 day supply of one or more statins	888 (270)	0.92 (0.79-1.07)
Murtola 2007 <sup>145</sup>	Prostate	Finland cancer registry	Finnish Cancer registry	Uncertain	Prescription database	1995-2002	Case-control	Non-Statin users	≥1 statin prescription	24723 (2622)	1.07 (1.00-1.16)
Khurana 2007 <sup>27</sup>	Lung	Eight states of South Central USA	VA database	Uncertain	Prescription database	1998-2004	Case-control	Non-Statin users	≥1 statin prescription	7280 (1994)	
Vinogradova 2007 <sup>30</sup>	Colorectal	UK general practices (QRESEARCH database)	GP Records	Invasive	GP Prescription records	1995-2005	Nested Case-control	Non-Statin users	≥1 statin prescription	5686 (538)	0.93 (0.83-1.04)
Setoguchi 2006 <sup>146</sup>	Colorectal Breast Lung	Elderly resident of Pennsylvania, ≥65 years with annual income <\$16,200)	Pennsylvania State Cancer Registry data	Invasive	Prescription database	1994-2003	Cohort	Glaucoma medication users	New user: ≥ 3 prescriptions during the first 180 days after the first prescription.	-*	No association*

[continued over]

[Table 3.3 continued]

Author & Year	Cancer(s) examined	Study Population	Data source for outcomes	Outcome definition	Data source for statin use	Time-period	Study design	Comparison group	Statin exposure definition	No. of cases (cases exposed)	Total cancer or site-specific cancer: point estimate (95% CI)
Friis 2004 <sup>147</sup>	Prostate Colorectal Breast Lung	Residents of the county of North Jutland, Denmark	Danish cancer registry	Uncertain	Prescription database	1989-2002	Cohort	Other lipid-lowering drugs	≥2 prescriptions during the study period - person time counted from the second prescription	22512 (398)	0.73 (0.55-0.98)
Graaf 2004 <sup>148</sup>	Prostate Colorectal Breast Lung	PHARMO record linkage system - 8 Dutch cities	Hospital linked discharge records	Invasive	Prescription database	1985-1998	Nested case-control	Cardiovascular drugs	≥6 months of statin use	3080 (144)	0.80 (0.66-0.96)
Kaye 2004 <sup>128</sup>	Prostate Colorectal Breast Lung	UK general practices (CPRD database)	GP Records	Uncertain	GP Prescription records		Nested case-control	Non-Statin users (untreated hyperlipidaemia)	≥1 prescription 1 year prior to diagnosis and statin use within a year of diagnosis	3244 (-)	1.0 (0.9-1.2)
Beck 2003 <sup>149</sup>	Breast	Saskatchewan health services database	Saskatchewan cancer registry	Uncertain	Prescription database	1989-1997	Cohort	Non-Statin users	≥1 statin at any time during the study period	879 (188)	1.09 (0.93-1.28)
Kaye 2002 <sup>129</sup>	Breast	UK general practices (CPRD database)	GP Records	Invasive and non-invasive	GP Prescription records	1992-1998	Nested case-control	Non-Statin users (untreated hyperlipidaemia)	≥1 prescription prior date of diagnosis	200 (31)	1.00 (0.6-1.6)
Blais 2000 <sup>150</sup>	Prostate Colorectal Breast Lung	10% random sample of individuals ≥65 from Regie de L'assurance-Maladie du Quebec (RAMQ) database	RAMQ Medical services records	Invasive	Prescription database	1988-1994	Nested case-control	Bile acid-binding resins	≥1 prescription at any time during the study period	65 (-)	0.72 (0.57-0.92)

GP: General Practice; CPRD: Clinical Practice Research Datalink; KP: Kaiser Permanente; VA: Veterans Affairs;

\*Hippisley-Cox *et al.*<sup>131</sup> and Setoguchi *et al.*<sup>146</sup> did not examine overall cancer, site-specific relative risk estimates are presented in **Table 3.4**



### 3.4.2.2 Outcome ascertainment

A variety of electronic data sources were used to ascertain cancer events (**Table 3.3**):

cancer registries (n=14), GP records (n=7), US administrative records (n=7), and

hospital records (n=2). Only one study, which utilised the VA database<sup>141</sup> compared

cancer outcomes ascertained from computerised records to medical charts in which

70% of cancer cases were verified. Vinogradova *et al.*<sup>90</sup> reported lower incidence rates

of colorectal cancer from the QRESEARCH database compared to 2003 ONS incidence

rates (rate difference: 6.2 (men), and 5.9 (women) per 100k PY).

### 3.4.2.3 Outcome definitions

Studies defined incident cancer by either including or excluding non-invasive

(carcinoma in-situ) cancers (**Table 3.3**). Of the 30 studies: 15 excluded non-invasive

cancers; while 4 studies included them. From reported case definitions, neither

inclusion nor exclusion of these cancer sub-types could be ascertained for the

remaining 11 studies. Statin users may be under closer monitoring compared to non-

users and therefore diagnosis of early detection of non-invasive cancers may be more

likely among statin users compared to their counterpart non-users.

### 3.4.2.4 Exposure ascertainment and definitions

Two main criteria were assessed by the 30 studies before classifying statin exposure

status (**Table 3.3**): (i) a minimum number of statin prescriptions written or dispensed;

and (ii) a minimum period of exposure. The most common requirement among 28 of

the 30 studies was a minimum of 1 statin prescription only (n=19);  $\geq 2$  statin

prescriptions only (n=4); and  $\geq 3$  statin prescriptions only (n=4).

Minimum periods of statin use or lag-time periods were considered by 18 studies,

which ranged from 3 months to 5 years:  $\geq 3$  months (~100 days' supply of statins; n=2);

≥6 months (n=3); ≥1 year (n=7); ≥2 years (n=5); and ≥5 years (n=1). Minimum periods of exposure may be considered for two reasons: firstly, to ensure only adherent statin users are included, and secondly, to guard against protopathic bias (or reverse causality), where patients are diagnosed with cancer after a short period of statin use, but had a pre-existing cancer prior to commencing statin therapy.

#### **3.4.2.4.1 New user design**

Four studies considered a new user design,<sup>56, 64, 131, 146</sup> only incident users of statins relative to the study period under consideration were included in the statin group.

When a *new user* design is implemented the exclusion of prevalent statin users may minimise prevalent user bias (**Prevalent user bias; Table 3.1**). Importantly, Farwell *et al.*<sup>25, 141</sup> noted that not all statin users could be confirmed as first time users within their study (proportions not given). Neither prevalent nor incident statin use was described in reported exposure methods; therefore the assumption was that both incident and prevalent statin users were included in both published analyses. Of note, all case-control studies did not specify whether statin users were incident or prevalent – the base cohort from which cases and controls are sampled could include only incident statin users as recommended by Ray *et al.*<sup>70</sup> This review assumed that the cohort from which case-control studies sampled from included both new and prevalent statin users.

#### **3.4.2.5 Comparison groups**

Five different comparator groups were considered (**Table 3.3**): the primary comparison group consisted of non-statin users (n=24); antihypertensive drug users (n=3);<sup>141, 148 132</sup> other lipid lowering drugs (n=3);<sup>128, 129, 147</sup> bile-acid binding resin drug users (n=1);<sup>150</sup> and glaucoma medication users (n=1).<sup>146</sup>

**Table 3.4: Summary of findings and biases - risk of breast, prostate, and colorectal and lung cancer associated with statin use**

Author + Year	Cancer	No. of cases (no. of cases in statin user group)	<sup>a</sup> Did the study potentially suffer from the following biases?					Time-window bias	Confounders adjusted for in analysis	Analysis method	Ever use: Relative risk (95% CI)
			Immortal time bias	Protopathic bias	Prevalent user bias	Healthy user bias					
<b>Cohort Studies</b>											
Farwell 2011 <sup>132</sup>	Prostate	546 (359)	No	No	Yes	No	-	Age, weight, co-morbidities, aspirin use, mental illness, alcoholism, lung disease, history of colonoscopy or sigmoidoscopy, smoking history, and total cholesterol.	Cox proportional hazards	HR=0.69 (0.52-0.90)	
Tan 2011 <sup>133</sup>	Prostate	2407 (565)	No	Yes	Yes	No	-	Age, BMI, race, DRE positivity, prostate volume, and number of cores surveyed.	Logistic Regression	OR=0.92 (0.85-0.98)	
Hippisley-Cox 2010 <sup>131</sup>	Breast Prostate Colon Lung	9,823 (-) 7,129 (-) Men: 2,182 (-) Women: 1,970 (-) Men: 3600 (-) Women: 2401(-)	U	Yes	No	Yes	-	<b>All cancer types:</b> Age, BMI, <b>Breast Cancer:</b> Townsend score, HRT, family history breast cancer, benign breast disease, oral contraceptive use, any other cancer. <b>Prostate Cancer:</b> smoking status. <b>Colorectal Cancer:</b> Townsend score, smoking status, colorectal polyps, type 2 diabetes. <b>Lung Cancer:</b> Townsend score, smoking status, any other cancer, corticosteroids, asthma	Cox proportional hazards	<b>Breast:</b> HR=1.09 (1.00 -1.18) <b>Prostate:</b> HR=1.05 (0.98 -1.13) <b>Colon:</b> Women HR= 0.89 (0.76- 1.05) <b>Colon:</b> Men HR= 0.89 (0.76-1.05) <b>Lung:</b> Women HR=1.10 (0.96-1.25) <b>Lung:</b> Men: HR=1.11 (1.01-1.23)	
Murtola 2010 <sup>134</sup>	Prostate	1,594 (268)	No	Yes	Yes	No	-	Age, family history of prostate cancer, use of aspirin, diabetic drugs, antihypertensives, no. of PSA screens, and calendar period of screening.	Cox proportional hazards with time-dependent exposure	HR=0.75 (0.63-0.89)	
Flick 2009 <sup>137</sup>	Colorectal	171 (56)	No	No	Yes	No	-	Family history of colorectal cancer, history of colorectal polyps, history of sigmoidoscopy, BMI, cardiovascular disease, hyperlipidaemia, physical activity, smoking, alcohol use, NSAID use, multivitamin use, red meat intake, calcium, folate intake, and ethnicity.	Cox proportional hazards	<b>Colorectal:</b> HR=0.89 (0.61-1.30) <b>Colon:</b> HR=0.90 (0.58-1.40); N=42 <b>Rectum:</b> HR=0.86 (0.41-1.78); N=14	
[Table 3.4 continued over]											

[Table 3.4 continued]											
Author + Year	Cancer	No. of cases (no. of cases in statin user group)	<sup>a</sup> Did the study potentially suffer from the following biases?					Time-window bias	Confounders adjusted for in analysis	Analysis method	Ever use: Relative risk (95% CI)
			Immortal time bias	Protopathic bias	Prevalent user bias	Healthy user bias					
Haukka 2009 <sup>67</sup>	Breast Prostate Colon Rectum Lung	6,046 (3048) 10,928 (5871) 2,950 (1486) 2,066 (1080) 5129 (2333)	U*	No	Yes	Yes	-	Age, sex, and follow-up period. <b>Note:</b> Also performed a sensitivity analysis for unmeasured confounders.	Poisson regression	<b>Breast:</b> IRR =1.01 (0.96-1.06) <b>Prostate:</b> IRR =1.12 (1.08-1.17) <b>Colon:</b> IRR =0.99 (0.92-1.06) <b>Rectum:</b> IRR =1.07 (0.98-1.17) <b>Lung:</b> IRR =0.81 (0.77-0.86)	
Singh 2009 <sup>64</sup>	Colorectal	6,637 (402)	Yes	Yes	No	No	-	Age, sex, history of IBD, diabetes, CHD, resective colorectal surgery, lower GI endoscopy, level of morbidity (three categories), SES, NSAID use, and HRT use.	Poisson regression	IRR: 1.13 (1.02-1.25)	
Boudreau 2008 <sup>139</sup>	Prostate	2,532 (246)	No	No	Yes	No	-	Age, diabetes, hypercholesterolemia, other lipid-lowering drug use, and NSAID use.	Cox proportional hazards with time-dependent exposure	HR=0.88 (0.76-1.02)	
Smeeth 2008 <sup>56</sup>	Breast Prostate	3,204 (324) 3,213 (312)	No	No	No	Yes	-	Age, sex, propensity score, year of index date, co-morbidities, co-medication	Cox proportional hazards	<b>Breast:</b> HR=1.17 (0.95-1.43) <b>Prostate:</b> HR=1.06 (0.86-1.30)	
Farwell 2008 <sup>141</sup>	Prostate Colorectal Lung	2,165 (1164) 687 (316) 867 (436)	No	No	Yes	No	-	Age, weight, co-morbidities, aspirin use, alcoholism, history of colonoscopy or sigmoidoscopy, smoking history, and total cholesterol.	Cox proportional hazards	<b>Prostate:</b> HR=0.90 (0.81-0.99) <b>Colorectal:</b> HR=0.65 (0.55-0.78) <b>Lung:</b> HR=0.70 (0.60-0.81)	
Friedman 2008 <sup>142</sup>	Breast Colorectal Lung	881 (-) Men:421 (-) Women: 312 (-) Men: 614 (-) Women: 482 (-)	No	No	Yes	Yes	-	Calendar year, hormone use, NSAID use (only for colorectal cancer). <b>Note:</b> Also performed external adjustment for smoking.	Cox proportional hazards with time-dependent exposure	<b>Women:</b> <b>Breast:</b> HR=0.99 (0.92-1.06) <b>Colon:</b> HR=0.97 (0.85-1.11) <b>Rectum:</b> HR=0.97 (0.76-1.25) <b>Lung:</b> HR=1.16 (1.06-1.28)  <b>Men:</b> <b>Colon:</b> HR=0.88 (0.78-1.00) <b>Rectum:</b> HR=0.93 (0.77-1.12) <b>Prostate:</b> HR=1.03 (0.98-1.08) <b>Lung:</b> HR=1.02 (0.94-1.11)	
	Prostate	1,706 (-)	No	No	Yes	No	-				

[Table 3.4 continued over]

[Table 3.4 continued]											
Author + Year	Cancer	No. of cases (no. of cases in statin user group)	<sup>a</sup> Did the study potentially suffer from the following biases?					Confounders adjusted for in analysis	Analysis method	Ever use: Relative risk (95% CI)	
			Immortal time bias	Protopathic bias	Prevalent user bias	Healthy user bias	Time-window bias				
Boudreau 2007 <sup>143</sup>	Breast	2,707 (130)	No	No	Yes	No	-	Age, HRT, diabetes, other lipid-lowering drugs, and BMI.	Cox proportional hazards with time-dependent exposure	HR=1.07 (0.88-1.29)	
Flick 2007 <sup>144</sup>	Prostate	888 (270)	No	No	Yes	No	-	Race, diabetes, KP California region.	Cox proportional hazards	HR=0.92 (0.79-1.07)	
Setoguchi 2006 <sup>146</sup>	Breast Colorectal Lung	300 (227) 233 (178) 216 (179)	No	No	No	No	-	Age, race, sex, health service utilisation, prevention-related activities, co-morbidities, GI drug use, HRT, NSAID use, Tobacco abuse.	Cox proportional hazards	<b>Breast:</b> HR=0.99 (0.74-1.33) <b>Colorectal:</b> HR=0.96 (0.70-1.31) <b>Lung:</b> HR=1.11 (0.77-1.60)	
Friis 2004 <sup>147</sup>	Breast Prostate Colorectal Lung	227 (48) 1,407 (34) 3,006 (55) 3399 (73)	U*	Yes	Yes	No	-	Age, gender, calendar period, NSAID use, HRT, cardiovascular drugs.	Poisson regression	<b>Breast:</b> IRR=1.02 (0.76-1.36) <b>Prostate:</b> IRR=0.87 (0.61-1.23) <b>Colorectal:</b> IRR=0.85 (0.65-1.11) <b>Lung:</b> IRR=0.92 (0.72-1.16)	
Beck 2003 <sup>149</sup>	Breast	879 (188)	No	Yes	Yes	Yes	-	Age, HRT, and oral contraceptive use.	Mantel-Haenszel relative risk	IRR=1.09 (0.93-1.28)	
<b>Case-Control studies</b>											
Chang 2011 <sup>68</sup>	Prostate	388 (83)	-	No	Yes	No	No	Sex, year of birth, index date, diabetes, hypertension, CHD, Benign prostatic hyperplasia, NSAID use, use of other lipid lowering drugs, number of physician visits, and number of hospitalisations	Conditional logistic regression	OR=1.55 (1.09-2.19)	
Vinogradova 2011 <sup>130</sup>	Breast Prostate Colorectal Lung	15,666 (1481) 14,764 (2774) 11,749 (2000) 10,163 (1998 <sup>67</sup> )	-	No	Yes	Yes	No	Age, Sex, Practice, Calendar time, co-morbidities, BMI, smoking status, Townsend deprivation score, family history of breast cancer, co-medications.	Conditional logistic regression	<b>Breast:</b> OR=1.00 (0.93-1.08) <b>Prostate:</b> OR=1.08 (1.01-1.14) <b>Colorectal:</b> OR=1.07 (1.00-1.15) <b>Lung:</b> OR=1.07 (0.99-1.16)	
Robertson 2010 <sup>135</sup>	Colorectal	9,979 (711)	-	No	Yes	Yes	No	Age, sex, NSAID use, diabetes, cholecystectomy, alcoholism, MI, stroke, Atherosclerosis	Conditional logistic regression	IRR =0.87 (0.80-0.96)	
Wooditschka 2010 <sup>136</sup>	Breast	22,488 (5409)	-	No	Yes	Yes	No	Oral contraceptive use and HRT use.	Unconditional logistic regression	OR=1.02 (0.97-1.08)	
Hachem 2008 <sup>138</sup>	Colorectal Rectal	6,080 (2,987)	-	Yes	Yes	No	No	Inflammatory bowel disease, diabetic nephropathy, colorectal evaluation, cholecystectomy, sulfonyleurea prescription, NSAID, and liver disease	Conditional logistic regression	OR=0.88 (0.83 – 0.93)	
[Table 3.4 continued over]											

[Table 3.4 continued]										
Author + Year	Cancer	No. of cases (no. of cases in statin user group)	<sup>a</sup> Did the study potentially suffer from the following biases?					Confounders adjusted for in analysis	Analysis method	Ever use: Relative risk (95% CI)
			Immortal time bias	Protopathic bias	Prevalent user bias	Healthy user bias	Time-window bias			
Boudreau 2008 <sup>140</sup>	Colorectal	357 (60)	-	Yes	Yes	Yes	No	Age, BMI, diabetes, smoking status, hormone therapy use, and NSAID use	Conditional logistic regression	Colorectal: OR=1.02 (0.65-1.59) Colon: OR=0.91 (0.55-1.50) Rectal: OR=1.47 (0.50-4.29)
Yang 2008 <sup>127</sup>	Colorectal	4,432 (-)	-	No	Yes	No	No	General practice site, calendar periods and duration of follow-up, BMI, Smoking status, alcohol use, HRT, NSAID use, colonoscopy, or flexible sigmoidoscopy.	Conditional logistic regression	≥5 yrs: OR=1.1 (0.5-2.2) ≥10 yrs: OR=1.3 (0.6-2.7)
Murtola 2007 <sup>145</sup>	Prostate	22,101 (2622)	-	Yes	Yes	Yes	No	Age, antihypertensives and diabetic medications	Conditional logistic regression	OR=1.07 (1.00-1.16)
Vinogradova 2007 <sup>90</sup>	Colorectal	5686 (538)	-	No	Yes	Yes	No	Smoking, obesity, deprivation, co-morbidities, use of the other medications	Conditional logistic regression	OR=0.93 (0.83-1.04)
Khurana 2007 <sup>27</sup>	Lung	7820	-	Yes	Yes	Yes	Yes	BMI, age, race, sex, alcohol use, diabetes	Logistic regression	OR=0.55 (0.52-0.59)
Graaf 2004 <sup>148</sup>	Breast Prostate Colorectal Lung	467 (-) 186 (-) 486 (-) 449 (-)	-	No	Yes	No	No	Diabetes mellitus, prior hospitalisations, chronic disease score, antihypertensives, hormone use, NSAIDs, other lipid-lowering therapy.	Conditional logistic regression	Breast: OR=1.07 (0.65-1.74) Prostate: OR=0.37 (0.11-1.25) Colon: OR=0.87 (0.48-1.57) Rectum: OR=0.48 (0.16-1.48) Lung: OR=0.89 (0.56-1.42)
Kaye 2004 <sup>128</sup>	Breast Prostate Colon Rectum Lung	40 (-) 62 (-) 25 (-) 23 (-) 43 (-)	-	Yes	Yes	No	No	Age, sex, general practice BMI, smoking status, GP visit frequency	Conditional logistic regression	Breast: OR=0.9 (0.6-1.3) Prostate: OR =1.3 (1.0-1.9) Colon: OR =1.0 (0.6-1.7) Rectum: OR =1.6 (0.9-2.8) Lung: OR =0.91 (0.6-1.3)
Kaye 2002 <sup>129</sup>	Breast	66 (41)	-	Yes	Yes	No	No	Hyperlipidaemia, HRT use, BMI, history of benign breast disease.	Conditional logistic regression	OR =1.0 (0.6-1.6)
Blais 2000 <sup>150</sup>	Breast Prostate Colon Lung	65 (-) 78 (-) 56 (-) 70 (-)	-	No	Yes	No	No	Age at index date, previous neoplasm, year of cohort entry, use of fibric acids, use of other lipid-lowering drugs and comorbidity score.	Logistic regression	Breast: IRR=0.67 (0.33-1.38) Prostate: IRR =0.74 (0.36-1.51) Colon: IRR =0.83 (0.37-1.89) Lung: IRR =0.94 (0.43-2.05)

<sup>a</sup> Detailed assessment of bias described in **Appendix B, Table 9.1**; \*U: Uncertain if study was potentially affected by potential biased based on lack of reported methods

BMI: Body Mass Index; CHD: Coronary Heart Disease; DRE: Digital Rectal Examination; GI: Gastro-Intestinal; GP: General Practice; HRT: Hormone Replacement Therapy; IBD: Inflammatory Bowel Disease; KP: Kaiser Permanent; MI: Myocardial Infarction; NSAID: Non-steroidal anti-inflammatory drug; PSA: Prostate-specific antigen; SES: Socio-economic status  
IRR: Incidence rate ratio; HR: Hazard Ratio; OR: Odds Ratio;

### 3.4.2.6 Adjustment for potential confounding factors

Adjustment for potential confounding factors was performed to varying degrees among the 30 studies (**Table 3.4**). Established risk factors such as age and sex were included in statistical adjustments by the majority of studies. However, adjustment for clinically relevant lifestyle factors and ethnicity were less frequent: smoking status (n=13), BMI (n=13); alcohol use (n=6); socioeconomic status (n=5); and ethnicity (n=5). This could be explained by studies utilising US claims databases, which do not collect lifestyle factors as part of routine practice. Co-medications (n=26) and co-morbidities (n=18) were included to varying degrees in statistical models. Three studies included propensity scores estimates as an adjustment factor within their statistical analysis.<sup>56, 132, 141</sup> A propensity score is a summary measure of how likely an individual is to be prescribed a particular drug. Use of a propensity score offers a method of reducing bias and confounding in pharmacoepidemiological studies.<sup>151</sup> Of note, Farwell *et al.*<sup>141</sup> found no difference in effect estimates on overall cancer risk when comparing risk estimates from propensity score adjusted analyses to traditional statistical adjustment.

### 3.4.3 Detailed consideration of site-specific associations

The risk of prostate cancer associated with statin use had the most variability among the four cancer types examined (**Figure 3.3**): 4 studies reported a reduced risk; 3 observed an increased risk; and 10 found no significant association. Variation in findings from studies examining the risk of colorectal cancer was also observed: 1 found an increased risk, 3 observed a reduced risk, and 13 reported no association. Fourteen studies examined the association between statin use and risk of breast cancer; all 14 studies reported a null association between statin use and breast cancer

risk. Eleven studies examined the risk of lung cancer associated with statin use. Of the 11 studies, 3 reported a reduced risk and 8 a null association.

### **3.4.3.1 Prostate cancer**

#### **Overview**

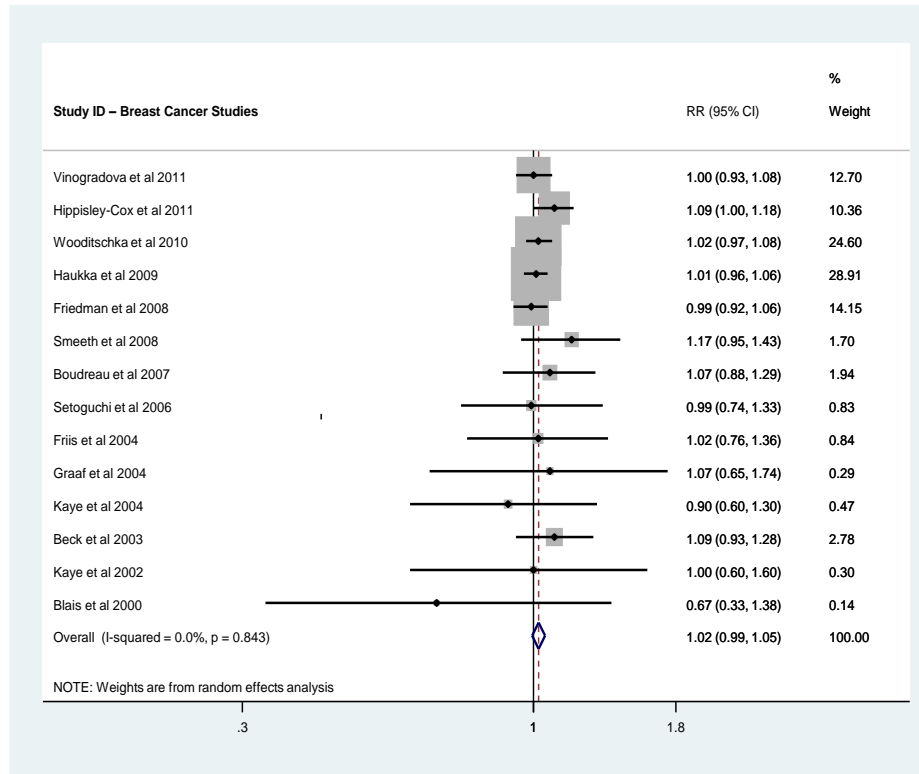
Seventeen studies examined the risk of prostate cancer associated with any statin use: 3 observed an increased risk; 4 reported a reduced risk; and 10 found no association (Table 3.2 and 3.4).

The majority of seventeen prostate cancer studies reported a null association (n=8) between statin use and the risk of prostate cancer. Point estimates varied among the 8 studies, the majority of the variation was from three small studies published between 2000-2004.<sup>128, 148, 150</sup> Relative risk (RR) point estimates ranged from 0.37<sup>150</sup> to 1.30.<sup>128</sup> Moreover, these three studies compared statin users to comparison drug groups, in contrast to most of the recent studies (post-2004) who compared statin users to non-users.



**Figure 3.3: Forest plots of studies examining statin use and breast, colorectal, lung, and prostate cancer risk**

**(a) Studies examining breast cancer risk associated with statin use**



**(b) Studies examining colorectal cancer risk associated with statin use**

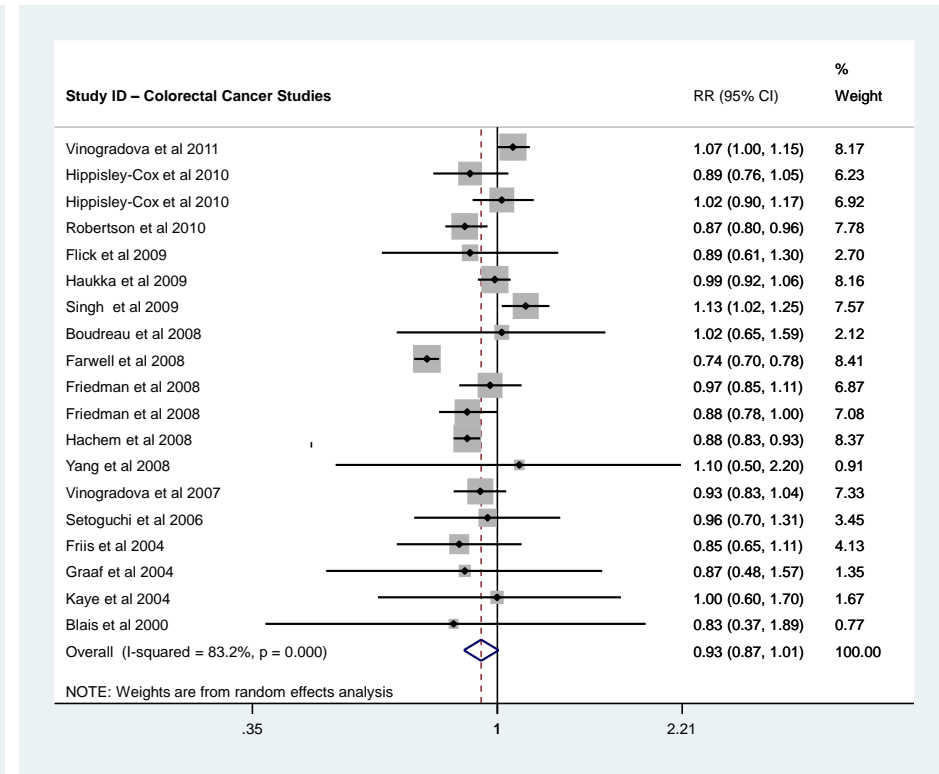
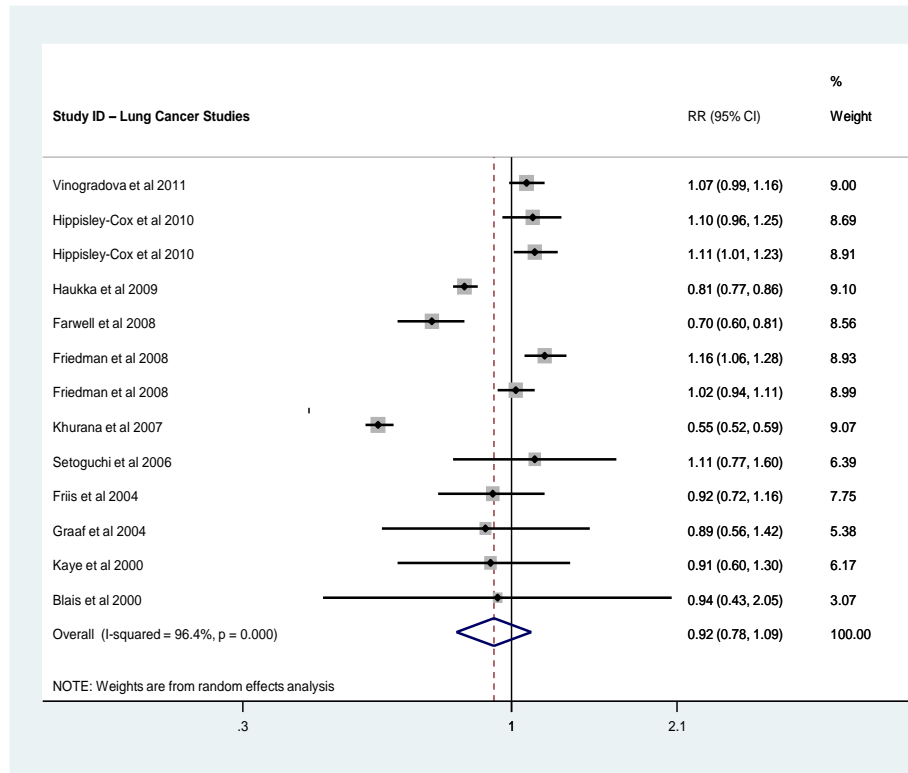
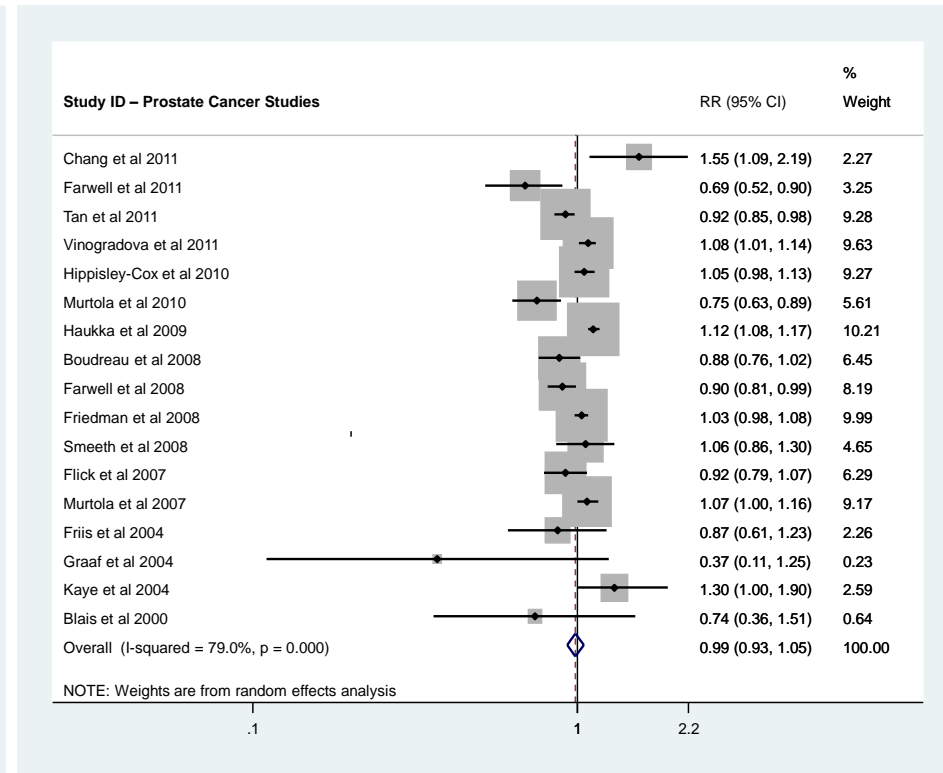


Figure 3.3 (continued): Forest plots of studies examining statin use and breast, colorectal, lung, and prostate cancer risk

(c) Studies examining lung cancer risk associated with statin use



(d) Studies examining prostate cancer risk associated with statin use



Two case-control studies and one cohort study observed an increased risk;<sup>68, 130</sup> in both cases, the lower bounds of reported confidence intervals were close to 1 (**Table 3.4**).

Chang *et al.*<sup>68</sup> reported a 55% increased risk of prostate cancer among statin users compared to non-users (OR=1.55; 95% CI, 1.09, 2.19). Cases were identified over a relatively short time-window (2005-2008) relative to the entire study period (1996-2008). In addition, a small number of cases (n=388) were included, generating relatively large confidence intervals. Of note, the pool of potential controls from the study conducted by Chang *et al.*<sup>68</sup> comprised of subjects who did not have a history of wrist or hip fractures - the same exclusions did not apply to potential cases.

Vinogradova *et al.*<sup>130</sup> reported a borderline increased risk (OR=1.08; 95% CI, 1.01, 1.14), in contrast to Chang *et al.*<sup>68</sup>, Vinogradova *et al.*<sup>130</sup> identified a fairly large number of cases of prostate cancer (n=14,764) over the study period (1998-2008). Haukka *et al.*<sup>67</sup> also reported an increased risk of prostate cancer (RR=1.12; 95% CI; 1.08, 1.17), however the study only adjusted for age, sex and duration of follow-up.

Of the seventeen prostate cancer studies, four cohort studies observed a reduced risk of prostate cancer associated with statin use. Two studies were conducted by Farwell *et al.*<sup>25, 141</sup> in the same population of VA veterans. In an attempt to minimise healthy user bias, a group of antihypertensive drug users were compared to statin users, which may have influenced the observed decrease in risk as there have been suggestions of an elevated risk of cancer among patients prescribed antihypertensives.<sup>152</sup> Tan *et al.*<sup>153</sup> conducted a cohort study consisting of men who had undergone consecutive prostate biopsies to circumvent healthy user bias. A reduced risk was observed (OR=0.92; 95% CI; 0.85, 0.98) among 565 prostate cancer cases (2407 cases overall) who had previously been prescribed statins. Similarly, a cohort study by Murtola *et al.*<sup>134</sup> also

observed a reduced risk (HR=0.75; 95% CI; 0.63, 0.89) sampled statin and non-statin users from the PSA screening arm of a Finnish clinical trial. All men from the trial had a PSA test every 4 years, which circumvented the issue of healthy user bias (differential PSA testing rates between treatment groups).

### **Duration of statin use**

Eight studies evaluated the relationship between duration of statin use and prostate cancer risk, three studies observed a reduced risk, four an increased risk, and one found no association (**Table 3.5**).

The short term effect of statin use were examined by four studies.<sup>67, 130, 134, 139</sup> A large nested case-control study conducted in the UK QRESEARCH database<sup>130</sup> observed a null association related to 13-24 months of statin use (OR=0.93; 95% CI; 0.81, 1.07) but this association did not hold when increasing duration to >25 months. Similarly, a study conducted in Finland<sup>67</sup> observed an increased risk of prostate cancer associated with less than 6 months (RR=1.27; 95% CI; 1.16, 1.38), however this relationship tended towards the null with 1 year of statin use. Furthermore, Murtola *et al.*<sup>134</sup> reported statistically significant effects associated with <3 years of statin use, although this association was much stronger within shorter periods of follow-up (1 year: HR=0.73; 95% CI; 0.54, 0.98; and 2-3 years: HR=0.67; 95% CI, 0.50, 0.90) compared to later durations of 4-5 years and ≥6 years (4-5 years; HR=0.85; 95% CI; 0.62, 1.11; ≥6 years: HR=0.70; 95% CI; 0.45, 1.08);  $p_{\text{trend}}=0.007$ ). Similar findings were observed by Boudreau *et al.*<sup>139</sup>, statin use for 1 - 2.9 years was associated with a reduced risk of prostate cancer (HR=0.75; 95% CI, 0.59, 0.95), however no association was found for 3-4.9 years (HR=0.92; 95% CI; 0.71, 1.19) and 5+ years (HR=1.06; 95% CI; 0.83, 1.34).

Three studies examined long-term statin use (>5 years).<sup>133, 142, 144</sup> Tan *et al.*<sup>133</sup> reported no association between long-term statin use and prostate cancer risk (OR=0.95; 95% CI; 0.79, 1.09) and similar findings were observed by Friedman *et al.*<sup>142</sup> In contrast, Flick *et al.*<sup>144</sup> examined whether NSAID use modified the effect of statins on the risk of prostate cancer because of previous evidence suggesting that NSAIDs have a protective effect among different cancers and hence may be considered a possible confounder when examining the association between statin use and cancer risk.<sup>90, 93, 154, 155</sup> Flick *et al.* observed a lower risk of prostate cancer among long-term statin users who were also regular NSAID users (HR=0.64; 95% CI; 0.44, 0.93). However, long-term statin use alone was not associated with prostate cancer (HR =1.05; 95% CI; 0.55, 1.98). Of note, a sensitivity analysis by Flick *et al.* compared analyses with and without a 1-year lag period; differences in results were negligible (For short term analysis: HR=0.94 with lag vs. OR=0.97 without lag and long-term analysis: 0.71 with lag vs. 0.72 without lag).

### 3.4.3.2 Colorectal cancer

#### Overview

Overall, 17 studies examined the association between statin use and risk of colorectal cancer: 1 found an increased risk, 3 observed a reduced risk, and 13 found no association (**Figure 3.3 (b)**). Among the 17 colorectal cancer studies, eight investigated colon and rectal cancer as separate outcomes (4/8 studies also examined colorectal cancer), two studies investigated colon cancer alone, and seven studies examined colorectal cancer only.

Of the 17 colorectal cancer studies, 13 reported null associations in relation to statin use. Point estimates ranged from 0.83<sup>150</sup> to 1.1.<sup>127</sup> Six of the thirteen studies were

underpowered yielding wide confidence intervals. Similar to prostate and breast cancer studies these were undertaken in the years 2000-2004 (**Figure 3.3 (b)**). However, several large studies were conducted post-2004 that observed a null association. A large nested case-control study utilising the QRESEARCH database<sup>130</sup> observed a null association (OR=1.07, 95% CI; 1.00, 1.15). Similar results were observed by Hippisley-Cox *et al.*<sup>131</sup> who also used the QRESEARCH database – primary results were stratified by sex and statin type. Haukka *et al.*<sup>67</sup> conducted a large cohort study among all Finnish residents and reported a null association between statin use and colon cancer risk (RR=0.99; 95% CI; 0.92, 1.06); however, only a limited number of potential confounding factors were included for adjustment (age, sex, and follow-up). A cohort study conducted in a population of residents from Manitoba, Canada<sup>64</sup> reported a borderline increased risk of colorectal cancer associated with statin use (RR=1.13; 95% CI; 1.02, 1.25). Of note, Singh *et al.*<sup>64</sup> observed statistically significant higher rates of lower gastro-intestinal endoscopy examinations among regular statin users compared to non-statin users: 28% vs 20%, similar differential rates were observed by Flick *et al.*<sup>137</sup> for sigmoidoscopies.

Three studies reported protective effects of statin use on colorectal cancer risk: two studies were conducted in the US<sup>138, 141</sup> and one study among two counties of Denmark.<sup>135</sup> A study utilising the VA healthcare database<sup>141</sup> found a reduced risk of colorectal cancer (HR=0.65; 95% CI; 0.55, 0.78) among statin users compared to antihypertensive drug users. Similarly, Hachem *et al.*<sup>138</sup> conducted a nested case-control study among diabetic veterans from the VA healthcare system and found a 12% reduction in colorectal cancer among statin users when compared to non-statin users (OR=0.88; 95% CI; 0.83, 0.93). Furthermore, Hachem *et al.*<sup>138</sup> stratified on

previous colon polyps and found no significant association in patients with previous polyps, however a protective effect was observed when analyses were restricted to patients without colon polyps (OR=0.86; 95% CI; 0.80, 0.93). Robertson *et al.*<sup>135</sup> examined the risk of colorectal cancer among statin users compared to non-users residing in two counties of Denmark and reported a 13% reduction in colorectal cancer risk in statin users compared to non-users (OR=0.87; 95% CI; 0.80, 0.96).

Of note, three studies utilised active comparator groups with contrasting findings. In contrast to Farwell *et al.*<sup>141</sup>, Graaf *et al.*<sup>148</sup> reported a null association (OR=0.87; 95% CI; 0.48, 1.57) of colon cancer risk when comparing statin users to antihypertensive drug users. However, based on the wide confidence interval reported the study lacked power. Setoguchi *et al.*<sup>146</sup> implemented an active comparator group of patients prescribed glaucoma medications and did not find any evidence of a difference in colorectal cancer incidence rates between the two groups (HR=0.96; 95% CI; 0.70, 1.31). Of note, both Setoguchi *et al.*<sup>146</sup> and Farwell *et al.*<sup>141</sup> observed comparable rates of colorectal examinations between the statin group and the respective comparator drug group.

### **Duration of statin use**

Six studies reported results relating to short term statin use (0-3 years)<sup>90, 130, 135, 138, 140, 146</sup> (**Table 3.5**). Three studies observed a reduced risk<sup>90, 135, 138</sup>. Vinogradova *et al.*<sup>90</sup> reported a marginal reduced risk associated with less than 1 year of statin use. Hachem *et al.*<sup>138</sup> reported a reduced risk associated with less than 6 months of statin use (OR=0.86; 95% CI; 0.77, 0.95), similarly Robertson *et al.*<sup>135</sup> also observed a reduced risk associated with 0-3 years of statin use (OR=0.84; 95% CI; 0.75, 0.95). Of note, when Hachem *et al.* restricted analyses to patients with no history of colorectal polyps

the protective association remained (OR=0.86; 95% CI; 0.75, 0.99). The remaining three studies did not observe an association between short term statin use and colorectal cancer risk.

Six studies assessed the effects of long term statin use on the risk of colorectal cancer:<sup>64, 127, 130, 135, 137, 142</sup> prolonged durations of >5 years were consistently shown to have no association with the risk of colorectal cancer among five of the six studies (**Table 3.5**). However, Vinogradova *et al.*<sup>130</sup> reported statistically significant increased risk associated with ≥49 months (~4 years) of statin use, and the corresponding test for trend was significant across stratified 1-year time periods over the 4-year period (p=0.002).

### 3.4.3.3 Breast cancer

#### Overview

Fourteen studies examined the association between statin use and risk of breast cancer; all 14 studies found no relationship between breast cancer risk and statin use. However, there were a range of point estimates reported (Relative risk range, 0.33<sup>150</sup> to RR=1.17<sup>56</sup>). Earlier studies reported wide confidence interval estimates around observed effect estimates demonstrating a lack of power. However, from 2004 onwards effect estimates were closer to the null with narrowing confidence intervals (**Figure 2.3 (a)**). Adjustment for potential confounders varied by study. The majority of studies adjusted for HRT (n=10), which may have increased the risk of breast cancer; in contrast, oral contraceptive use was only included for adjustment by 4 studies.

#### Duration of statin use

Six studies investigated the effects of statin duration on breast cancer risk (**Table 3.5**).

Only one study reported an increased risk associated with short term statin use (<6



months). Short durations were investigated by 4 studies.<sup>130, 143, 146, 149</sup> Beck *et al.*<sup>149</sup>

observed an increase in breast cancer risk associated with <6 months of statin use.

Setoguchi *et al.*<sup>146</sup> and Vinogradova *et al.*<sup>130</sup> did not observe any significant effects on breast cancer associated with <3 years and <2 years respectively.

Similarly, long-term durations of statin use did not show any significant associations in five studies.<sup>130, 142, 143, 146, 149</sup> Friedman *et al.* looked at >5 years of statin use (HR=1.06; 95% CI; 0.86, 1.21); estimates from the other four studies were similar (**Table 3.5**).

#### **3.4.3.4 Lung cancer**

##### **Overview**

Overall, eleven studies examined lung cancer risk associated with statin use, two reporting an increased risk, and three a decreased risk, and six a null association (**Table 3.2**).

Two studies reported an increased risk of lung cancer (**Table 3.4, Figure 3.3**).<sup>131, 142</sup>

Friedman *et al.* observed an elevated risk among women (HR=1.16; 95% CI, 1.06, 1.28).

In contrast, Hippisley-Cox *et al.*<sup>131</sup> observed a marginal increased risk among men (HR=1.11; 95% CI; 1.01, 1.23). In both cases, results were borderline and may have been a result of unmeasured confounding or chance findings.

Of the 11 studies, 3 reported a reduced risk (**Table 3.4**). Among the three studies,<sup>27, 67,</sup>

<sup>141</sup> none adjusted for smoking status – Farwell *et al.*<sup>141</sup> extracted smoking status;

however, over 50% of statin users and >60% of antihypertensive drug users (referent group) had an unknown smoking status. Although 6 studies observed a null association, most of these studies were underpowered, reporting relatively large confidence intervals and adjustment for smoking status varied by study. The exception

being Vinogradova *et al.* a large case-control study conducted in the QRESEARCH database, adjusting for several important confounding factors including smoking status. Vinogradova *et al.*<sup>63</sup> reported a null association (OR=1.07; 95% CI; 0.99, 1.16).

### **Duration of statin use**

Three studies examined the risk of cancer associated with short term statin use (0-3years). Vinogradova *et al.*<sup>63</sup> did not find a significant association between short term statin use. Similar findings were reported by Setoguchi *et al.*<sup>146</sup> (<3 years: HR=1.18; 95% CI; 0.72, 1.92). In contrast, Khurana *et al.*<sup>27</sup> reported an elevated risk associated with less than 6 months of statin use (OR=2.32; 95% CI; 2.05, 2.63). However, an inverse association was reported with increasing duration of statin use >6 months (1-2 years: OR=0.70; 95% CI; 0.61, 0.79).

Four studies examined the risk of lung cancer associated with long term statin use, two of which observed a null association.<sup>142, 146</sup> In contrast, Vinogradova *et al.*<sup>63</sup> reported an increased risk (>4 years: OR=1.18; 95% CI; 1.05, 1.34), while Khurana *et al.*<sup>27</sup> reported a reduced risk of cancer (>4 years: OR=0.23; 95% CI; 0.20, 0.26).

### **3.4.4 Assessment of bias**

For each study, summarised assessments of five pre-selected biases are presented in **Table 3.4**; detailed assessments of each bias are provided in **Appendix B, Table 9.1**.

Overall, many of the studies included in this review may have been potentially affected by at least one of the biases considered: 1/16 cohort studies by immortal time bias; 26/30 by prevalent user bias; 10/30 by healthy user bias; 1/14 case-control studies by time-window bias; and 12/30 from protopathic bias. Varied methodology and populations between the studies made it difficult to un-tangle specific bias effects.

#### 3.4.4.1 Immortal time bias

Only one of the 16 cohort studies potentially suffered from immortal time bias (**Table 3.4**). Singh *et al.*<sup>64</sup> reported a borderline increased risk of colorectal cancer associated with statin use, statin use was defined as a minimum of two statin prescription during follow-up. However, observation for cancer events began at the first statin prescription rather than second potentially inducing immortal time bias. Immortal time bias would have biased the rate ratio downward, suggesting an underestimate of the risk reported. That being said, protopathic bias could have also affected findings of this study. Hippisley-Cox *et al.*<sup>156</sup> and Friis *et al.*<sup>147</sup> were also identified as studies potentially affected by immortal time bias; however, based on described methods from both studies, immortal time bias could not be assessed with certainty. The remaining 13 cohort studies avoided immortal time bias by incorporating either a time-dependent exposure of statin use or excluded immortal time bias.

#### 3.4.4.2 Protopathic bias

Overall, 12 studies did not implement a lag period or minimum period of exposure, which could have made them susceptible to protopathic bias (**Table 3.4**). Of the 12 studies, 4 examined short term statin use ( $\leq 12$  months) associated with cancer risk: 2 observed an increased risk of cancer (1 lung cancer<sup>27</sup> and 1 breast cancer<sup>149</sup>), and 2 a reduced risk of cancer (1 colorectal cancer<sup>138</sup> and 1 prostate cancer<sup>29</sup>). Relative risk estimates incorporating all periods of follow-up time were similar for the two studies initially observing a reduction in cancer risk.<sup>29, 138</sup> However, the 2 studies initially reporting an increased risk of lung and breast cancer associated with short term statin use observed a reduced risk and null association when examining all time periods of follow-up respectively.

In comparison, 18 studies implemented a lag-period, of which 5 examined short term statin use ( $\leq 12$  months) associated with cancer risk: 4 observed a null association<sup>90, 127, 130, 137</sup> and 1 an increased risk of prostate cancer.<sup>67</sup> Short term statin-cancer associations observed remained for 4 of the 5 studies when examining all periods of statin follow-up. Vinogradova *et al.*<sup>130</sup> observed a borderline increased risk of prostate cancer when considering all follow-up time (OR=1.08; 95% CI; 1.01-1.14). Of note, none of the 9 studies that examined short term statin use associated with cancer risk implemented a new user design.

#### **3.4.4.3 Prevalent user bias**

The majority of studies included in this review (26/30) included prevalent statin users within their study. Four cohort studies restricted exposure to incident statin use,<sup>56, 64, 131, 146, 156</sup> 3 utilised a comparator group of non-statin users, while one compared *new* statin users to glaucoma medication users.<sup>146</sup> Of the 4 studies, 3 reported a null association between statin use and cancer risk. Of note, Singh *et al.*<sup>64</sup> performed two sets of exposure analyses: (i) examining the dose effect of statin use among new and prevalent statin users; and (ii) the same dose-response analysis restricted to *new statin* users. Interestingly, compared to the *any user* analysis, *new user* risk estimates were marginally reduced (rate ratio=1.35 any use vs 1.28) when low doses ( $\leq 1$  daily defined dose, DDD) were considered; however, risk estimates (rate ratio = 0.79 any vs 0.82 new) increased when high doses ( $> 1$  DDD) were examined. Marginally higher relative risk point estimates were reported by Smeeth *et al.*<sup>56</sup> and Hippisley-Cox *et al.*<sup>131</sup> in comparison to their counterpart studies (cohort design of *any* statin users and non-users) (**Figure 3.3**), suggesting a marginal downward biasing effect from the inclusion of prevalent statin users.

#### 3.4.4.4 Healthy user bias

Eight studies compared statin users to an active comparator group to minimise healthy user bias (**Section 3.4.2.5**). Setoguchi *et al.*<sup>146</sup> observed a null association when comparing statin users to glaucoma medication users. Of note, physician visits at treatment start date were similar between treatment groups (mean no of physician visits – statin users 9.1 vs glaucoma medication users 9.6). Similarly, null associations were observed from studies comparing statin users to non-statin lipid lowering drugs<sup>128, 129, 147</sup> and bile-acid binding drugs.<sup>150</sup> However, most of these studies were generally underpowered due to the high prevalence of statin use and low number of patients in the respective comparator groups. Farwell *et al.*<sup>25, 141</sup> reported a reduced risk of colorectal, lung, and prostate cancer among statin users compared to antihypertensive drug users (**Table 3.4**). Although there is some evidence of an association between antihypertensive drug use and cancer risk, which may have influenced findings.<sup>96, 157</sup>

The majority of studies in this review compared statin users to non-users; however, some studies (n=12) adjusted for potential factors that could contribute to healthy user bias; for example, PSA testing, GP/physician visits, no. of hospitalisations, or cancer related testing. Of particular note is the adjustment of PSA testing in prostate cancer studies. Larger proportions of statin users have been shown to have had a PSA test compared to non-users, which could increase detection of prostate cancer.

Murtola *et al.*<sup>134, 145</sup> conducted two studies, one adjusting for PSA testing, the other not. The study that did not adjust for PSA testing reported a null association (1.07; 95% CI; 1.00-1.16) between statin use and prostate cancer risk.<sup>145</sup> In contrast, once

adjusting for PSA testing, a reduced risk of prostate cancer was observed (HR=0.75; 95% CI, 0.63, 0.89).

Studies accounting for healthy user bias brought about the most variability in terms of relative risk estimates compared to studies that apparently did not adjust for healthy user bias. However, this may be due to a lack of power in studies utilising comparator groups or an association existing between the comparator drug group and the cancer of interest, or unmeasured confounding.

#### **3.4.4.5 Time-window bias**

Only one of the 14 case-control studies was potentially affected by time-window bias.<sup>27</sup> Khurana *et al.*<sup>27</sup> sampled controls from their last observation point during study follow-up; which potentially overestimated the proportion of exposed controls relative to exposed cases (statin exposure – controls (34%), cases (27%)). Of note, no lag period was implemented which may have made the study susceptible to protopathic bias. The remaining 13 case-control studies minimised time-window bias by either incorporating risk-set sampling of controls or matching controls on case date of diagnosis. However, 12 of the remaining 13 studies reported higher relative risk estimates (except Graaf *et al.*<sup>148</sup> – although an alternative drug group of antihypertensive drug users were considered as a comparator group) compared to Khurana *et al.*<sup>27</sup> suggesting a downward biasing effect of time-window bias.

#### **3.4.5 Meta-analysis**

Four main meta-analyses (random effects models) were undertaken for each cancer type. All 95% confidence intervals for the summary relative risk estimates spanned 1 (Figure 3.3). Among the breast cancer studies there was low heterogeneity ( $I^2=0.0\%$ ,

p=0.0843); however, high heterogeneity was observed for colorectal ( $I^2=83.2\%$ , p=0.000), lung ( $I^2=96.4\%$ , p=0.000), and prostate cancer ( $I^2=79.0\%$ , p=0.000).

Subgroup analyses were conducted to examine the stability of the pooled relative risk estimates. Choice of any user or new user as the exposure group yielded the most variability: among studies of breast cancer that focussed on new users vs any users, the SRR was 1.09, 95% CI (1.01, 1.18); for colorectal cancer any user vs non-user yielded SRR=0.91, 95% CI (0.38, 0.99), and for lung cancer, new user vs non-user yielded SRR=1.11, 95% CI (1.02, 1.20). Other comparisons with non-users all resulted in non-significant effects for breast, colorectal, and prostate cancer; however, a pooled protective effect of lung cancer among statin users compared to active comparison drug groups was observed SRR=0.7, 95% CI(0.71, 0.79). There was no variation by geographical location in the associations between statin use and breast, colorectal, or lung cancer, but a non-significant protective effect of statin use was observed for North American studies examining prostate cancer SRR=0.93, 95% CI(0.86, 1.00).

**Table 3.5: Association between duration of statin use and cancer risk – secondary analysis results**

	Breast Cancer		Colorectal Cancer		Lung Cancer		Prostate Cancer	
Author & Year	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users
Chang 2011 <sup>68</sup>	Cancer type not investigated	-	Cancer type not investigated	N/A	Cancer type not investigated	N/A	0-5 years: 0.92 (0.85-0.99) >5 years: 0.95 (0.79-1.09)	495 70
Farwell 2011 <sup>132</sup>	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A	No analysis	-
Tan 2011 <sup>133</sup>	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A	No analysis	-
Vinogradova 2011 <sup>130</sup>	<12 months: 1.01 (0.90-1.13) 13-24 months: 0.93 (0.81-1.07) 25-48 months: 1.07 (0.95-1.21) ≥49 months: 0.95 (0.83-1.09) Ptrend=0.719	433 289 430 329	<12 months: 1.05 (0.95-1.17) 13-24 months: 1.04 (0.92-1.17) 25-48 months: 1.02 (0.92-1.14) ≥49 months: 1.23 (1.10-1.38) Ptrend=0.002	525 400 539 536	<12 months: 1.02 (0.90-1.15) 13-24 months: 1.11 (0.97-1.27) 25-48 months: 1.01 (0.90-1.14) ≥49 months: 1.18 (1.05-1.34) Ptrend=0.013	485 406 549 558	<12 months: 1.05 (0.95-1.15) 13-24 months: 1.14 (1.03-1.26) 25-48 months: 1.09 (0.99-1.19) ≥49 months: 1.05 (0.95-1.16) Ptrend=0.084	668 560 796 750
Murtola 2010 <sup>134</sup>	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A	1 year : 0.73 (0.54-0.98) 2-3 years: 0.67 (0.50-0.90) 4-5 years: 0.85 (0.62-1.11) ≥6 years: 0.70 (0.45-1.08) Ptrend=0.007	60 95 60 53
Robertson 2010 <sup>135</sup>	Cancer type not investigated	N/A	0-3 years: 0.84 (0.75-0.95) 3-5 years: 0.88 (0.74-1.04) >5 years: 0.95 (0.80-1.12)	370 162 179	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Wooditschka 2010 <sup>136</sup>	No analysis	-	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Hippisley-Cox 2010 <sup>131</sup>	No analysis	-	No analysis	-	No analysis	-	No analysis	-
Flick 2009 <sup>137</sup>	Cancer type not investigated	N/A	Colorectal cancer: 101 days<5 years: 0.91 (0.61-1.34) ≥5 years: 0.83 (0.43-1.63)	45 11	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Haukka 2009 <sup>67</sup>	No analysis	-	No analysis	-	No analysis	-	<0.5 year: 1.27 (1.16-1.38) 1 year: 0.98 (0.97-1.00)	Unspecified

[Table 3.5 continued over]



[Table 3.5 continued]

	Breast Cancer		Colorectal Cancer		Lung Cancer		Prostate Cancer	
Author & Year	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users
Singh 2009 <sup>64</sup>	Cancer type not investigated	N/A	≥5 years (New & prevalent users) <1.14 DDD: 1.01 (0.75-1.37) ≥1.14 DDD: 0.75 (0.51-1.10)  ≥5 years (New users) <1.18 DDD: 0.90 (0.54-1.49) ≥1.18 DDD: 0.69 (0.37-1.28)	43 27  15 10	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Boudreau 2008 <sup>140</sup>	Cancer type not investigated	N/A	Colorectal: <2 years: 0.80 (0.40-1.59) ≥2 years: 1.22 (0.70-2.12)	20 40	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Boudreau 2008 <sup>139</sup>	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A	1 - <3 years: 0.75 (0.59-0.95) 3 - <5 years: 0.92 (0.71-1.19) ≥5 years: 1.06 (0.83-1.34)	79 66 87
Farwell 2008 <sup>141</sup>	Cancer type not investigated	N/A	No analysis	-	No analysis	-	No analysis	-
Friedman 2008 <sup>142</sup>	>5 years statin use 1.02 (0.86-1.21)	136	>5 years statin use Women: 1.02 (0.75-1.38) Men: 1.00 (0.78-1.30)	42 62	>5 years statin use Women: 1.17 (0.93-1.46) Men: 1.06 (0.88-1.28)	78 119	>5 years statin use 1.04 (0.93-1.17)	322
Hachem 2008 <sup>138</sup>	Cancer type not investigated	N/A	<6 months: 0.86 (0.77-0.95) 6-12 months: 0.98 (0.89-1.09) 12-18 months: 0.91 (0.82-1.02) 18-24 months: 0.93 (0.84-1.04) >24 months: 0.87 (0.80-0.95)	144 225 180 201 366	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Smeeth 2008 <sup>56</sup>	No association: data not shown	Unspecified	Cancer type not investigated	N/A	Cancer type not investigated	N/A	No association: data not shown	Unspecified
Yang 2008 <sup>127</sup>	Cancer type not investigated	N/A	5 years: 1.1 (0.8-1.6) 6 years: 1.2 (0.8-1.8) 7 years: 1.2 (0.7-2.0) 8 years: 1.2 (0.7-2.2) 9 years: 1.3 (0.7-2.4) 10 years: 1.3 (0.6-2.7)	Unspecified Unspecified Unspecified Unspecified Unspecified	Cancer type not investigated	N/A	Cancer type not investigated	N/A

[Table 3.5 continued over]

[Table 3.5 continued]

	Breast Cancer		Colorectal Cancer		Lung Cancer		Prostate Cancer	
Author & Year	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users	Duration, point estimate (95 % CI)	No of cases: statin users
Boudreau 2007 <sup>143</sup>	1 - <3: 0.96 (0.71-1.31) 3 - <5: 1.04 (0.72-1.51) ≥5: 1.27 (0.89-1.81) Ptrend=0.2	~4541 ~30 ~35	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Flick 2007 <sup>144</sup>	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A	101 days - <5 yrs: 0.97 (0.83-1.13) ≥5 years: 0.72 (0.53-0.99)	228 42
Murtola 2007 <sup>145</sup>	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A	No analysis	-
Khurana 2007	Cancer type not investigated	N/A	Cancer type not investigated	N/A	0-0.5 yrs: 2.32 (2.05-2.63) 0.5-1.0 yrs: 0.75 (0.63-0.89) 1.0-2.0 yrs: 0.70 (0.61-0.79) 2.0-4.0 yrs: 0.49 (0.44-0.55) >4.0 yrs: 0.23 (0.20-0.26)	446 214 416 649 269	Cancer type not investigated	N/A
Vinogradova 2007 <sup>90</sup>	Cancer type not investigated	N/A	1-12 months: 0.84 (0.71-1.00) 13-24 months: 0.99 (0.80-1.22) >24 months: 0.99 (0.84-1.16)	183 122 233	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Setoguchi 2006 <sup>146</sup>	<3 years: 0.66 (0.42-1.05) ≥3 years: 1.28 (0.90-1.84)	47 156	<3 years: 0.93 (0.57-1.51) ≥3 years: 0.97 (0.65-1.46)	74 104	<3 years: 1.18 (0.72-1.92) ≥3 years: 1.02 (0.59-1.74)	99 80	Cancer type not investigated	N/A
Friis 2004 <sup>147</sup>	No analysis	-	No analysis	-	No analysis	-	No analysis	-
Graaf 2004 <sup>148</sup>	No analysis	-	No analysis	-	No analysis	-	No analysis	-
Kaye 2004 <sup>128</sup>	No analysis	-	No analysis	-	No analysis	-	No analysis	-
Beck 2003 <sup>149</sup>	1-90 days: 1.59 (1.14-2.20) 91-180 days: 2.02 (1.42-2.89) 181-365 days: 0.92 (0.57-1.49) 1-2 years: 1.48 (1.06-2.07) 2-3 years: 1.07 (0.68-1.69) 3-4 years: 0.95 (0.55-1.65) ≥4 years: 0.26 (0.12-0.55)	38 32 17 37 19 13 7	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Kaye 2002 <sup>129</sup>	No analysis	-	Cancer type not investigated	N/A	Cancer type not investigated	N/A	Cancer type not investigated	N/A
Blais 2000 <sup>150</sup>	No analysis	-	No analysis	-	No analysis	-	No analysis	-

## **3.5 Discussion**

### **3.5.1 Overview**

A total of 30 relevant studies were identified for this review. Overall, most early studies (pre 2004) observed a null association between statin use and the risk of breast, colorectal, lung, and prostate cancer; but most lacked power to detect a true association. Importantly, recent and large observational studies reported conflicting findings, particularly for the risk of colorectal, lung, and prostate cancer among statin users. Among the 30 studies, there was variability in terms of applied design methods, adjustment for potential confounders, as well as range of differing populations in which the studies were undertaken.

### **3.5.2 Meta-analysis**

The meta-analysis of 30 observational studies examining the risk of breast, colorectal, lung, and prostate cancer among patients prescribe statins support no evidence of an association. Studies examining colorectal, lung, and prostate cancer showed the most variability in relative risk estimates. In subgroup analyses, new users of statins compared to non-users showed marginal increased risks of breast and lung cancer. A protective effect of lung cancer was observed when comparing statin users to active drug comparator groups. Studies conducted among North American patients yielded a non-significant protective effect of prostate cancer among statin users compared to non-users. Although subgroup analyses showed significant pooled effects among studies implementing a new user design or an alternative drug comparator group, their effects may have been intertwined (4 new user design studies in which 3 incorporated a comparator drug group). In addition, studies incorporating comparison drug groups were underpowered due to the high

prevalence of statin use (other lipid lowering drugs and glaucoma medication drugs).

### **3.5.3 Assessment of bias**

Among the 30 studies, several common biases were examined which could explain part of the variation in findings between observational studies examining the risk of cancer associated with statin use. The biases included immortal time, protopathic, prevalent user, healthy user, and time-window bias. Prevalence of the different biases varied between the 30 studies: 1/16 cohort studies may have potentially suffered from immortal time bias; 12/30 from protopathic bias; 26/30 from prevalent user bias; 10/30 from healthy user bias; and 1/14 cases-control studies from time-window bias.

#### **3.5.3.1 Immortal time bias**

Immortal time bias could have potentially affected findings reported by one cohort study. However, the impact of immortal time bias may have been minimal due to the conservative exposure definition (statin user: minimum of 2 statin prescriptions) implemented by Singh *et al*<sup>64</sup>. That being said, the study conducted by Singh *et al.* might have also been affected by protopathic bias which could have influenced the rate ratio upward or downward. In the former situation, this may have diluted or negated any effects of immortal time bias.

#### **3.5.3.2 Protopathic bias**

In general, most cancer types typically have a long latency period. Statin exposure shortly before diagnosis may have little effect on the development of cancer. In comparison to studies that did implement a lag period minimising protopathic bias, short term risk estimates from studies that did not implement varied in terms of

direction (2 reduced and 2 increased risk of cancer). Importantly, although studies examined the risk of cancer among short term statin users, a new user analysis was not implemented in any of the study designs, making it difficult to assess the relationship between statin duration and cancer risk and also potentially inducing prevalent user bias.

There is no general consensus on how to define exposure optimally to firstly mitigate protopathic bias, and secondly to ensure a cohort of adherent statin users.<sup>158-161</sup> However, there are several important issues that need to be considered before defining exposure status. Firstly, studies investigating more than one cancer type implemented minimum periods of exposure in a global fashion. Optimal lag-periods may vary depending on the drug and cancer type under investigation.<sup>116, 130</sup> Secondly, time related selection bias must be kept in mind when decisions about periods of follow-up between treatment groups are made.<sup>69, 119</sup> Lastly, the adoption of minimum periods of exposure may be more appropriate when implemented alongside a new user design, rather than implemented arbitrarily among prevalent users, where cohort entry is a time point unrelated to exposure. This allows for the duration of statin use to be investigated with greater precision.

### **3.5.3.3 Prevalent user bias**

In relation to the previous subsection, new user designs were implemented by four studies; however, all studies utilised a prescription database which could have made it easier to implement a new user design.<sup>56, 64, 131, 132, 141, 146</sup> All new user studies reported a null association between statin use and the risk of any of the cancer types of interest. The effect of utilising a new user design was examined by only one study.<sup>64</sup> Singh *et al.*<sup>64</sup> performed two sets of dose-response analyses on

colorectal cancer risk: with and without prevalent statin users. Interestingly, when comparing the two analyses the exclusion of prevalent users reduced the relative risk of colorectal cancer at low doses and drove the estimate upward at high doses. Higher relative risk estimates of cancer were observed from the two cohort studies that compared *new* statin users to non-users compared to cohort studies that utilised a prevalent statin cohort. This protective effect is consistent with prevalent user bias discussed by Ray *et al.*<sup>70</sup> However, caution is needed in making this assertion due to competing biases and potential unmeasured confounding from the observational studies included in this review.

The new user design has two practical benefits: it allows for the mitigation of prevalent user bias and depletion of susceptible patients; as well as accurate assessment of statin duration. However, the utilisation of an inception cohort with a comparable drug group may be problematic due to the high prevalence of statin use among elderly patients. For example, the study by Setoguchi *et al.*<sup>146</sup> was underpowered, when considering a comparison drug group of new glaucoma medication users. In addition, statins are the primary drug of choice when treating hyperlipidaemia: the prevalence of other lipid lowering drugs such as bile acid-binding resins, and fibrates have declined since statins were first licensed in the early 1990s.

#### **3.5.3.4 Healthy user bias**

It has been shown that compared with the general population, statin users have good adherence to medications. In addition, they may also receive more counselling regarding a healthy lifestyle. Both of these factors are predictors of good health which may lead to an increased detection or prevention of early

cancers.<sup>162</sup> Furthermore, statin users may also be more likely to receive or be adherent to cancer screening,<sup>117</sup> for example PSA testing, and therefore may have higher cancer incidence rates compared to the untested general population. These healthy user and detection biases can be mitigated by implementing a new user design and/or comparing statin users to a group with similar healthy characteristics.

### **Comparison groups**

The majority of the reviewed studies compared statin users to non-statin users when assessing cancer risk; four studies compared statin users to other drug groups: antihypertensive drug groups; glaucoma medication users; and other non-statin lipid lowering drugs. Farwell *et al.*<sup>132, 141</sup> reported a reduced risk of prostate, and colorectal cancer among statin users compared to men prescribed antihypertensive medications. However, there is some evidence of an increased risk of cancer associated with antihypertensive drug use,<sup>157, 163</sup> hence the selection of antihypertensive drug users as a comparator group may lead to a spurious reduced risk among statin users. Setoguchi *et al.*<sup>146</sup> compared *new statin* users to new users of glaucoma medications. However, no significant effects of statin use associated with colorectal or breast cancer were observed when compared to glaucoma medication users. Importantly, there have been no previous reports of an association between glaucoma medications and cancer risk.

Non-statin lipid-lowering drug groups were utilised by two studies.<sup>147, 150</sup> Although non-statin lipid-lowering drug users are a relatively rare group due to the increasing number of patients prescribed statins in preference over non-statin drugs.

Characteristics of statin and non-statin lipid lowering drugs may be similar in terms of healthy behaviour. Both studies found a significant reduction in risk of overall

cancer, however when individual cancers were examined non-significant reduced risks were observed.

The utilisation of a non-statin group from the general population has been considered a limitation by some.<sup>123, 146</sup> However, the utilisation of comparators from another drug-class has several disadvantages: (i) different co-morbidities from statin users may influence results; (ii) the comparator drug chosen may be associated with the disease of interest, and (iii) there may be a possible reduction in precision due to a smaller number of patients available in the comparator group.

### **Adjustment for health utilisation services**

Several studies included in this review showed that statin users have a higher likelihood of PSA testing compared to non-statin users from the general population.<sup>139, 144, 145</sup> This could explain the increased risk of prostate cancer found in four of the studies included in this review,<sup>67, 68, 130, 145</sup> - none of which avoided potential detection bias by PSA testing. Conversely, there is evidence suggesting that statin use is associated with decreased PSA levels,<sup>139, 164</sup> hence PSA testing may lead to missed cases of prostate cancer and a false reduced risk associated with statin use.<sup>165</sup> All the studies in this review that observed a reduced risk of prostate cancer associated with any statin use also attempted to minimize detection bias by either study design or through a subset analysis. However, the rate of PSA testing has been shown to vary by country and healthcare system:<sup>166</sup> in parts of the US, the rate of PSA testing among men is higher compared to that of the UK, where no recommendations for prostate cancer screening have been made. This would imply that the impact of detection bias via PSA testing may also differ from country to



country – and would possibly be less of a biasing factor in UK studies compared to studies conducted in the US.

#### **3.5.3.5 Time-window bias**

Based on reported methods, one of the 14 case-control studies included in this review was susceptible to time-window bias.<sup>27</sup> Khurana *et al.*<sup>27</sup> reported a 45% reduction of lung cancer risk associated with statin use compared to non-users. Controls were sampled independently of time from their last point of contact, leading to a higher proportion of exposed controls due to their longer treatment observation period. Suissa *et al.*<sup>35, 119</sup> identified this potentially biased study and others that may have also been affected by time-window bias. The other 12 case-control studies avoided time-window bias by either matching controls to case date of diagnosis or risk-set sampling. Notably, all 12 case-control studies comparing statin users to non-users reported overall higher relative risk estimates compared to the potentially biased study conducted by Khurana *et al.*,<sup>27</sup> suggesting a consistent effect of time-window bias.

### 3.5.4 Updated review studies

An update of the systematic review was conducted to examine the current findings from studies undertaken after the original systematic review. Overall, ten studies were identified from a re-run of the original literature search in July 2015

**(Appendix Table B, Table 9.2).**

Overall, 2 studies examined the effects of statin use on breast cancer risk,<sup>167, 168</sup> 5 on colorectal cancer risk,<sup>168-172</sup> 2 on lung cancer risk,<sup>168, 173</sup> and 4 on prostate cancer risk<sup>168, 174-176</sup> – one study examined all four cancer types<sup>168</sup> **(Appendix C, Table 9.3).**

Both breast cancer studies observed a null association. For colorectal cancer, four studies observed a null association. In contrast, one study observed a reduced risk (HR=0.84; 95% CI; 0.76, 0.92) of colorectal cancer when comparing *new statin* users to new glaucoma medication users. Two studies reported no association of lung cancer risk associated with statin use compared to non-users. Two prostate cancer studies reported a reduced risk, while the remaining study reported an increased risk (OR=1.24; 95% CI; 1.10, 1.42).

Some of the studies may have been potentially affected by at least one bias: 1/10 cohort studies by immortal time bias; 8/10 by prevalent user bias; 5/10 by Healthy user bias; 1(uncertain)/10 case-control studies by time-window bias; and 6/9 from protopathic bias **(Appendix B, Table 9.3).**

Only two studies examined duration of statin use associated with cancer risk, both for prostate cancer. Jespersen *et al.*<sup>174</sup> reported null associations for 0-4 years of statin use, and a marginal reduction in risk for statin use between 5-9 years (OR=0.93; 95% CI; 0.88, 0.98). In contrast, Lustman *et al.*<sup>175</sup> reported consistent

reduced risks of prostate cancer associated with short (0-12 months: HR=0.68; 95% CI; 0.60, 0.79) and long term statin use (5+ years: HR=0.22; 95% CI; 0.17, 0.26).

Findings from the updated review studies were generally consistent with the original systematic review: (i) variation in reported relative risk estimates by cancer type; (ii) wide range of populations and design methods utilised; and (iii) the low prevalence of immortal and time-window bias and relatively high prevalence of protopathic, prevalent user, and healthy user bias.

### **3.5.5 Future research**

There are a number of suggestions for future research. Overall, only four studies examined the effects of statin use on cancer risk using a new user design, and two of the six studies conducted analysis relating to statin dose. Further research into the effects of new user designs in combination with dose-response analysis would be helpful in assessing prevalent user bias. Additionally, similar comparison groups relative to the exposure groups could also be identified, although this may be difficult due to the high prevalence of statin use among many populations.

Furthermore, the assessment of optimal lag times among different cancer types could be addressed in more detail. In the majority of cases, lag times were based on the investigators' subject knowledge and are not necessarily consistent between studies or transparently justified.

### **3.6 Conclusion**

Overall, there was no strong evidence of an association between statin use and risk of breast, colorectal, lung, or prostate cancer from the reviewed observational studies. However, there were conflicting findings from some studies, which might have been affected by potential biases. Whether these findings were driven by

particular biases is difficult to ascertain due to the varying methodology applied, various adjustment for potential confounder, and differing populations from which these studies were undertaken. The impact of the potential biases considered in this review will be further examined in the context of the statin-cancer association, **Chapter 6.**

### **3.7 Summary**

- A systematic review of statin use and the risk of breast, colorectal, lung, and prostate cancer among studies utilising routinely collected electronic patient records was conducted with an emphasis on methodological considerations.
- 30 studies were identified, 14 case-control and 16 cohort studies. Studies were conducted in a variety of populations and various routinely collected electronic health data sources.
- All studies were assessed for potential biases including: immortal time, protopathic, prevalent use, healthy user, and time-window bias.
- Overall, there were conflicting findings between studies, particularly for cancers of the colorectum, lung, and prostate.
- Prevalence of immortal time and time window-bias was low. However, greater than a third of all studies were potentially affected by either protopathic, prevalent user, or healthy user bias.
- The impact of biases examined in this review was difficult to assess due to competing biases, different populations, and varying methodology between studies. However, there was some evidence of healthy user, prevalent user and time-window bias influencing overall findings.

## **4 Data Sources and Methods**

### **4.1 Introduction**

This chapter describes the main data sources utilised and methods applied to both main analysis chapters (**Chapters 5 and 6**) of this thesis.

### **4.2 Data Sources**

#### **4.2.1 The Clinical Practice Research Datalink (CPRD)**

##### **4.2.1.1 Overview**

The Clinical Practice Research Datalink is a primary care database containing anonymised patient records from computer systems used by general practitioners in the UK. The CPRD began data collection from general practices in 1987.

Currently, the CPRD hold medical records for about 8% of the UK population with around 12 million patient records.<sup>177, 178</sup>

A typical dataset provided by the CPRD (CPRD GOLD, GP OnLine Database) contains patient information such as date of birth, sex, and details of registration with the practice. In addition, longitudinal data on clinically relevant lifestyle factors such as body mass index (BMI), smoking status, and alcohol status are also recorded.

Detailed information on prescriptions, clinical events, specialist referrals, and hospital admissions are also recorded. General practitioners enter all the medical information through a computerised system which codes clinical events using a Read coding system.<sup>177</sup>

Read codes are the standard hierarchical classification system used to record medical information in UK primary care settings. They were specifically developed for use in primary care by Dr James Read during the 1980s. There are approximately

250,000 Read codes used to record patient diagnoses, symptoms, and processes of care (e.g. referrals to secondary care).<sup>179</sup> Currently, the CPRD uses the 5-byte Read (or Read version 2) dictionary to code medical events; codes are continually added to the dictionary over time, though never removed. Each Read code is linked to a specific phrase of text, which indicates a single diagnosis or symptom. Diagnostic codes start with a letter whereas symptoms, signs, investigations, procedures and administration tasks start with a number. Previously, the Oxford Management Information Systems (OXMIS) coding system was used by the CPRD to code all clinical data. All practices transferred to the Read dictionary at varying dates in the 1990s.<sup>180</sup> Once data have been collected from the general practices, the CPRD perform a series of assessments at both practice and patient level to ensure a high quality standard of data.<sup>177</sup>

#### **4.2.1.2 Data access and extraction**

Studies that plan to access data from the CPRD require approval from the Independent Scientific Advisory Committee for MHRA database research (ISAC) (**Appendix C.1**).

The LSHTM hold flat data files (CPRD GOLD datasets) provided by the CPRD (updated every 6 months). The flat files were processed and formatted (Stata format) by the Electronic Health Records (EHR) group at LSHTM; bespoke data extraction from the flat files is available on ISAC approval. All CPRD GOLD data used for purposes of this thesis were extracted from the July 2012 version of the CPRD. In order to address **Aims 1-4** of this thesis (**Chapter 1, Section 1.7**), several samples of the CPRD database were used to conduct analyses, detailed methods are described in each chapter (**Chapter 5, Section 5.3 and Chapter 6, Section 6.3**). Case

identification methods applicable to **Chapters 5 and 6** are described in **Sections**

**4.2.2-4.2.4**. Further case identification methods specific to each main analysis are described in each respective chapter.

#### **4.2.1.3 Included events**

Four specific cancer types were elected as primary events for this study: breast, colorectal, lung, and prostate. These were selected because of their importance to pharmacoepidemiological research and because they among the most common cancer types diagnosed in men and women in the UK.<sup>181</sup>

#### **4.2.2 Cancer diagnostic groups**

A cancer diagnosis code list used in past studies<sup>30, 97, 102</sup> to identify cancer outcomes was modified for the purposes of this study (**Appendix C, Table 10.1-10.5**). Cancer diagnosis codes were classified into six groups: (1) malignant neoplasms; (2) in-situ cancers; (3) history of cancer; (4) borderline (uncertain whether malignant or benign); (5) suspected (suspected cancer, abnormal screening test, or fast track referral); and (6) general malignant neoplasms (site unspecified). In addition, Read codes were mapped to equivalent International Statistical Classification of Diseases and Related Health Problems, 10<sup>th</sup> Revision (ICD-10) diagnosis codes in order to be consistent with case definitions used by the ONS. All codes were reviewed and classified by KB, MR, and LS; any disagreements were reviewed again until resolved.

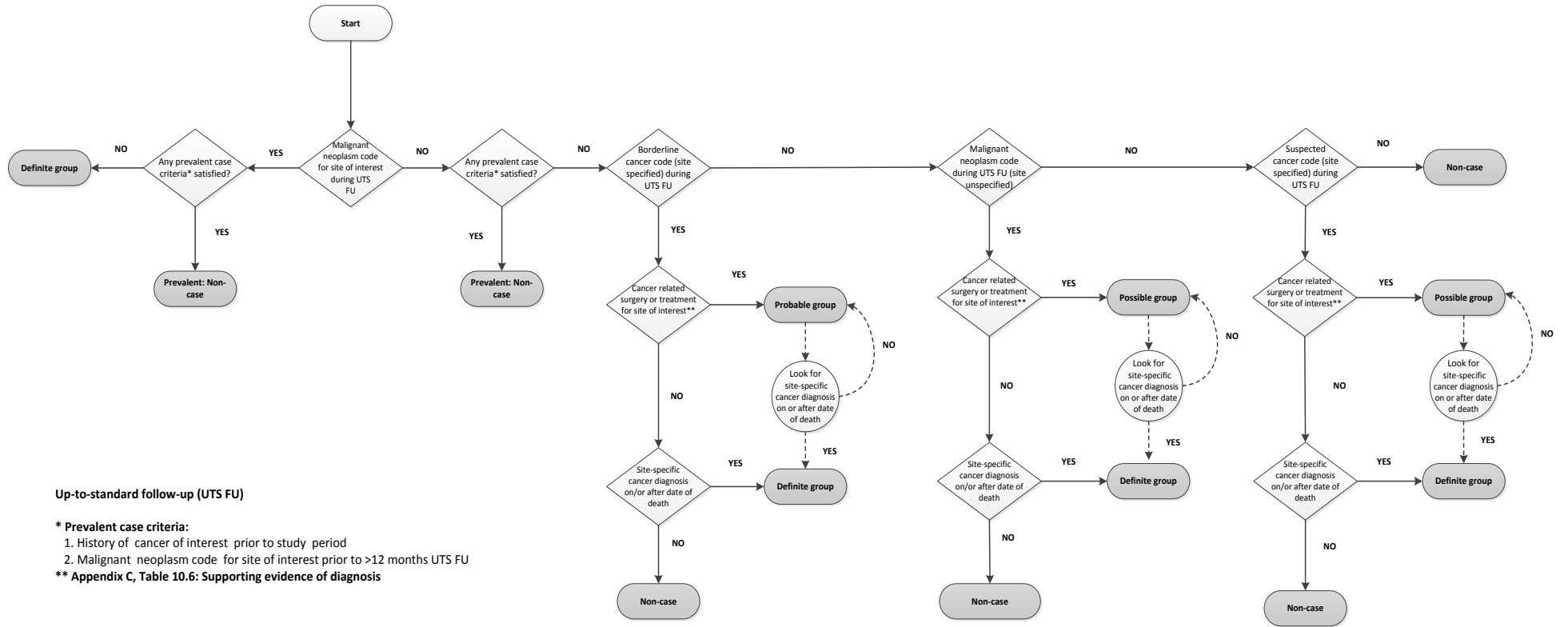
Three diagnostic groups were developed; **Figure 4.1** depicts the algorithm used to classify patients into certain groups. A hierarchical approach was used. First, patients with a recorded malignant diagnosis were grouped into the “definite” diagnostic group. Second, further searches of patient records grouped patients into a “probable” diagnostic group if they had a borderline diagnosis code. Lastly,

patients were grouped into a “possible” group if they had a suspected or general diagnosis code recorded within their computerised records.

Evidence of diagnosis such as cancer related surgery, chemotherapy, visits to an oncologist assisted in confirming potential cases that did not have a recorded malignant diagnosis code within their patient records. Evidence of diagnosis used as supporting evidence of diagnosis is listed in **Appendix C, Table 10.6.**



**Figure 4.1: CPRD diagnostic algorithm**



### 4.2.3 Cancer case definitions

Two case definitions were implemented (**Table 4.1**):

- (i) (i) A “standard” definition included patients from the definite diagnostic group. At least one definite diagnosis of a cancer of interest was required;
- (ii) (ii) A “broad” definition included patients from all three diagnostic groups (i.e. definite, probable, and possible).

These definitions were developed for the following reasons: (i) they reflect how researchers typically defined cancer in UK primary care databases (**Chapter 2**); (ii) inclusion of cases with non-specific diagnoses, such as borderline or suspected diagnoses, may have been misclassified in the “standard” definition as non-cases, which may influence primary care estimated incidence rates (**Chapter 2**); (iii) earlier non-specific diagnoses may be detected differentially between patients prescribed a preventative medication compared to non-users from the general population.

**Table 4.1: Case definitions**

<b>Case definition</b>	<b>Diagnostic group*</b>	<b>Description</b>
<b>Standard</b>	<b>Definite</b>	≥1 site-specific malignant neoplasm diagnosis code corresponding to ICD-10: C18-C20 (colorectum); C34 (lung); C50 (breast); C61( prostate)
<b>Broad: Definite + probable + possible</b>	<b>Definite</b>	≥1 malignant neoplasm diagnosis code corresponding to ICD-10: C18-C20 (colorectum); C34 (lung); C50 (breast); C61( prostate)
	<b>Probable</b>	≥1 borderline diagnosis code with supportive evidence of diagnosis (site known) during UTS follow-up
	<b>Possible</b>	≥1 Malignant neoplasm diagnosis code (site unspecified) with supportive evidence of diagnosis (site known) during UTS follow-up  ≥1 suspected diagnosis code and further evidence of diagnosis

Detailed algorithm depicting the classification of “definite”, “probable”, and “possible” diagnostic groups is given in **Figure 4.1**  
 ICD: International Classification of Diseases; UTS: Up-to-standard

#### 4.2.4 Externally linked data sources

Two external data sources were linked to the CPRD: cancer registry and ONS mortality. Overview of the individual data sources and linkage coverage are described in the following sections.

##### 4.2.4.1 National Cancer Data Repository (NCDR)

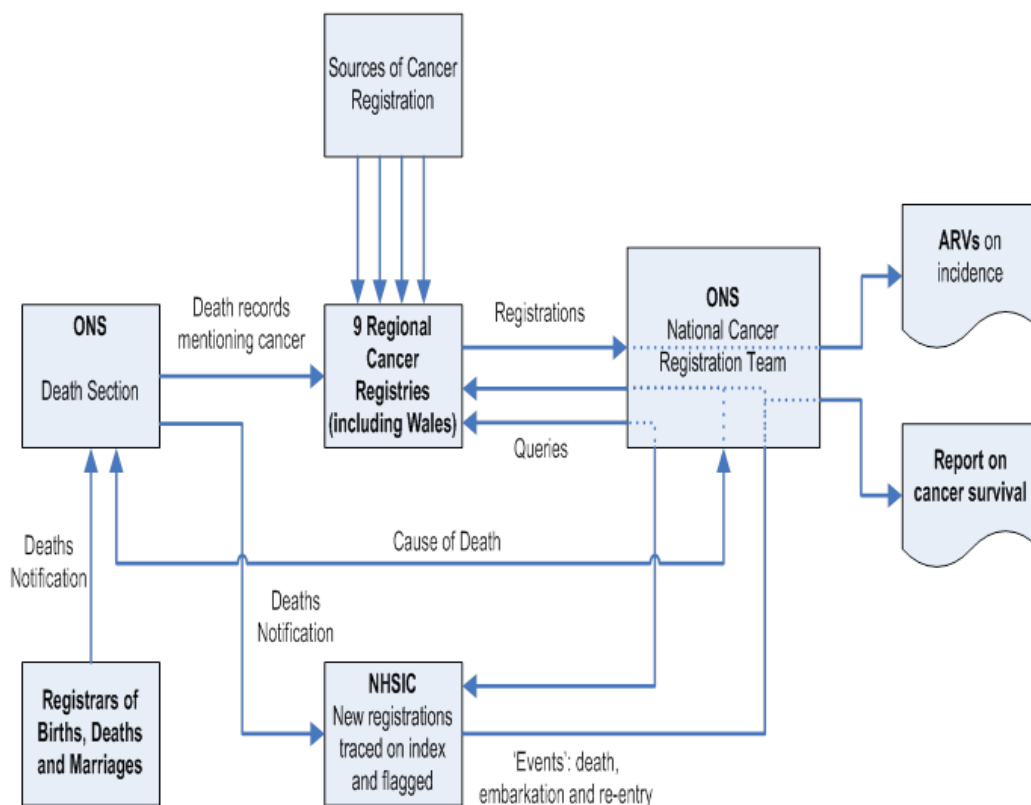
In England, there are 8 regional population based cancer registries; English cancer registrations are collated by the NCDR and are individually linked to primary care records for patients in the CPRD linkage scheme.<sup>178, 182</sup> Since 1993, it has been mandatory for the NHS, including trusts, to provide key items on demographics and diagnosis to regional cancer registries. Clinical information on *new* cancer registrations typically include date of diagnosis, site, death certificate indicator (indicating if a diagnosis was confirmed by death certificate only), treatments, stage and grade of diagnosis, and diagnosis by a screening process.<sup>182</sup>

Cancer registrations held by the 8 regional registries are received from a range of NHS sources (e.g. GPs, hospitals, coroners, radiotherapy) and are centrally stored by the ONS cancer registration computer system. Cases with either duplicated records or true multiple primary records are linked together by a probability matching process and are queried back to the regional registries until resolved **(Figure 4.2)**.<sup>182</sup>

Once all data have been received by the ONS and checks have been made, the ONS compiles detailed statistical tables on mortality, survival, prevalence, and incidence rates of different cancer types by age, sex, and geographical location.<sup>182</sup> The ICD coding dictionary is currently adopted by the ONS to code all cancer registrations. The ICD classifies diseases into broad groups known as chapters, and its worldwide

use enables numbers of deaths from different causes to be compared both between countries and over time. The World Health Organisation (WHO) has coordinated these revisions for many years.

**Figure 4.2: Flow diagram illustrating sources of cancer registrations [reproduced from Cancer Statistics registrations Series MB1<sup>182</sup>]**



#### 4.2.5 ONS Mortality (Death Registry)

In the UK, the ONS collects information on cause of death from civil registration records. Since 1837, it has been mandatory for all deaths in the UK to be registered.<sup>183</sup> The legal requirement to certify and register all deaths occurring in England and Wales means that death registrations provide the most complete data source for mortality statistics.

Currently, the ONS holds two databases on death related data: (i) a registration database which contains textual information derived from the death certificate;

and (ii) a statistical database which contains coded details on each death.<sup>183</sup> Once on the ONS database, data are passed through a series of automated validation processes which highlight any inconsistencies. Validation checks include consistency between dates of birth, death and registration, as well as consistency between date of death and employment status. Inconsistencies are checked through regular contact with the identified registrars to resolve any issues identified.<sup>183</sup> Cause of death is coded using the ICD-9/10. For the majority of deaths (around 80%), ONS codes the underlying cause of death using automated cause coding software. Remaining deaths are processed manually by experienced coders, mainly from deaths that have received a coroner's inquest.<sup>183</sup> A typical death registry dataset includes date of death, underlying cause of death, and other contributing diseases or conditions leading to the underlying cause of death.

#### **4.2.6 Linkage to the CPRD: cancer and death registry**

Patients in the CPRD were linked to the NCDR and death registry (ONS Mortality) data. The linkage was carried out in December 2014 by a trusted third party (The Health and Social Care Information Centre, HSCIC), using a deterministic match between NHS number, date of birth, and sex.<sup>184</sup> Currently, only a subset of English practices are part of the linkage scheme, some individual patients are excluded due to opting out or lack of a valid NHS identifier. NCDR and death registry linkage cover approximately 70% of the contributing CPRD English practices, or about 55% of contributing CPRD UK practices.

All analyses that incorporated linked data were restricted to patients who were eligible and consented to participate in the linkage scheme. Set 9 of the CPRD linked dataset was used, which covered the following time periods: **Cancer registry,**

from January 1, 1990 to December 31, 2010; and **ONS death registrations**, from January 1, 1998 to January 10, 2012.

Deaths registered after January 1, 2001 were coded using ICD-10. Prior to this date, ICD-9 had been in use across the UK since 1979. All ICD-9 codes were mapped to ICD-10 to accommodate the overall coding used in this thesis.

When patient data was linked by the HSCIC, several variables were used to judge the certainty of patient-level linkage, such as NHS number, date of birth, and sex. Each CPRD patient linked to ONS mortality data was given a “match rank” score based on how many variables were used to link patient records.<sup>185</sup> For analyses presented in this thesis (**Chapter 5 and 6**), if multiple death dates were observed for a patient, records with the highest match rank were retained. When a patient had >1 record with the same “match rank” score, the earliest date of death record was used for all analyses.

## **5 Validity of cancer diagnosis in the CPRD: comparison of observed and expected cancer incidence rates and concordance with national cancer registrations**

### **5.1 Introduction**

This chapter describes a series of analyses comparing CPRD estimated cancer incidence rates to UK ONS reported rates. In addition, this chapter also measures the concordance of recorded cancer diagnoses between the CPRD (primary care), NCDR (cancer registry), and ONS mortality (death registry).

### **5.2 Objectives**

For the four most common cancers, breast, colorectal, lung, and prostate cancer, this chapter aimed to:

1. Compare incidence rates calculated from primary care data (CPRD) with published national incidence rates based on cancer registrations (ONS).
2. Measure concordance of recorded diagnoses between the CPRD, NCDR, and ONS mortality.
3. Assess the impact of incorporating cancer registry data linked to the CPRD when estimating incidence rates.

### **5.3 Methods**

#### **5.3.1 Patients**

For computational reasons, a random sample of 2 million eligible patients was

selected from the CPRD. Eligible patients had to meet the following criteria:

acceptable record flag (e.g. consistent recording of age, sex, registration details and event recording);<sup>177</sup> subjects with acceptable registration status (e.g. permanently registered with a practice; no out of sequence year of birth or registration date; no



missing or invalid transfer out date; valid and non-missing year of birth; no missing sex information). In addition, subjects were excluded if they had a diagnosis or history of the cancer of interest prior to the start of follow-up.

### **5.3.2 Outcomes**

Four specific cancers were elected as outcomes for this study: breast, colorectal, lung, and prostate. These were selected because of their importance to pharmacoepidemiological research and for being the most common cancers diagnosed among men and women in the UK.<sup>186</sup>

### **5.3.3 Follow-up time**

The start of follow-up for each subject was the latest of the following dates: January 1, 2000, or 6 months after UTS registration with a practice. The end of follow-up was defined as the earliest of the following: December 31, 2010 (end of study period), a diagnosis of the cancer of interest; the date that the patient transferred out of the practice; date of death, or the last date for data collection by the practice. Follow-up was limited to the end of 2010 because published ONS incidence rates were only available up to 2010 at the time of conducting this study.

### **5.3.4 CPRD diagnostic groups and case definitions**

Diagnostic groups and case definitions for the four cancer types examined in this thesis are described in **Chapter 4, Section 4.2.2 and 4.2.3.**

### **5.3.5 Concordance of recorded cancer diagnoses between primary care, linked cancer registry, and death registry data**

Details of the linkage are described in **Chapter 4, Section 4.2.4**. Cancer diagnoses in the NCDR and ONS death data were coded using ICD-10. Read codes were mapped to ICD-10 to enable consistency of recorded diagnoses between the CPRD and linked data sources (**Chapter 4, Section 4.2.2**). If a patient had an incident diagnosis of breast (C50), colorectal (C18, C19, C20), lung (C34), or prostate (C61) cancer in the CPRD and also had a corresponding diagnosis (same ICD-10 code) in the NCDR, then this would be classified as a concordant diagnosis.

However, if a diagnosis of either of the cancers of interest was recorded in the NCDR and not in the CPRD, then a number of pre-determined factors were investigated to shed light on the discrepancy. In particular, the following were described: **(i)** any cancer diagnoses in the CPRD (e.g. different or non-specific cancer types), **(ii)** related cancer diagnoses in the CPRD (same site: in situ, borderline, suspected); **(iii)** NCDR age at diagnosis; **(iv)** time from NCDR diagnosis to CPRD defined end date **(v)** death recorded in the CPRD.

In the other direction, if a diagnosis of any of the cancer types was recorded in the CPRD and not in the NCDR, then information on any malignant diagnoses in the NCDR was sought. Pre-determined factors included: **(i)** frequency of recorded diagnosis in the CPRD, **(ii)** time from CPRD recorded diagnosis to death, **(iii)** age at CPRD diagnosis, and **(iv)** time from registration to CPRD diagnosis (to assess inclusion of prevalent cases).

### **5.3.6 Linked ONS death certificates and cancer registry incidence rates**

Three combinations of linkages were considered when calculating incidence rates:

(i) CPRD and ONS death registrations; (ii) CPRD and cancer registry data; and (iii)

CPRD, ONS death registrations and cancer registry. For each of the combinations,

two estimates were calculated: (1) incidence rates among the linkage population,

but restricted to CPRD data only, and (2) supplemented rates from (1) incorporating

additional cases from the respective linked data source (NCDR and/or ONS

mortality).

### **5.3.7 Statistical analysis**

#### **5.3.7.1 Denominator**

Individuals were eligible to contribute to the denominator population if their

records were UTS and they had been followed on the database for  $\geq 6$  months

before the beginning of the year of interest. For each age and sex category, the

denominator person-time was estimated by summing the number of days each

eligible individual contributed to each category, by calendar year and dividing this

sum by 365.25.

#### **5.3.7.2 Incidence rate estimates**

CPRD incidence rates initially calculated using primary care data only, were

compared to reported UK ONS rates.<sup>182</sup> For each cancer type, crude age- and sex-

specific incidence rates were estimated for each year over the study period with

corresponding 95% confidence intervals according to a Poisson distribution. Age

was categorised into 5-year age groups through to 80 years of age, and then a

single age-group for 85 years and older.

#### 5.3.7.4 Age-standardised incidence rate estimates

Directly age-standardised incidence rates (ASIR) were estimated using the European Standard Population<sup>182</sup> given by:

$$ASIR = \{\sum AIR_k P_k\} / \sum P_k, \quad K= 0, 1-4, 5-9, \dots, 80-84, \text{ and } 85 \text{ and over}$$

$AIR_k$  = observed incidence rate in age group k

$P_k$  = European standard population in age group k

Corresponding 95% confidence intervals were calculated using the following formula:

$$ASIR \pm ASIR / (\sqrt{\sum n_k}), \text{ where } n_k \text{ is the number of events observed in sex/age group k}$$

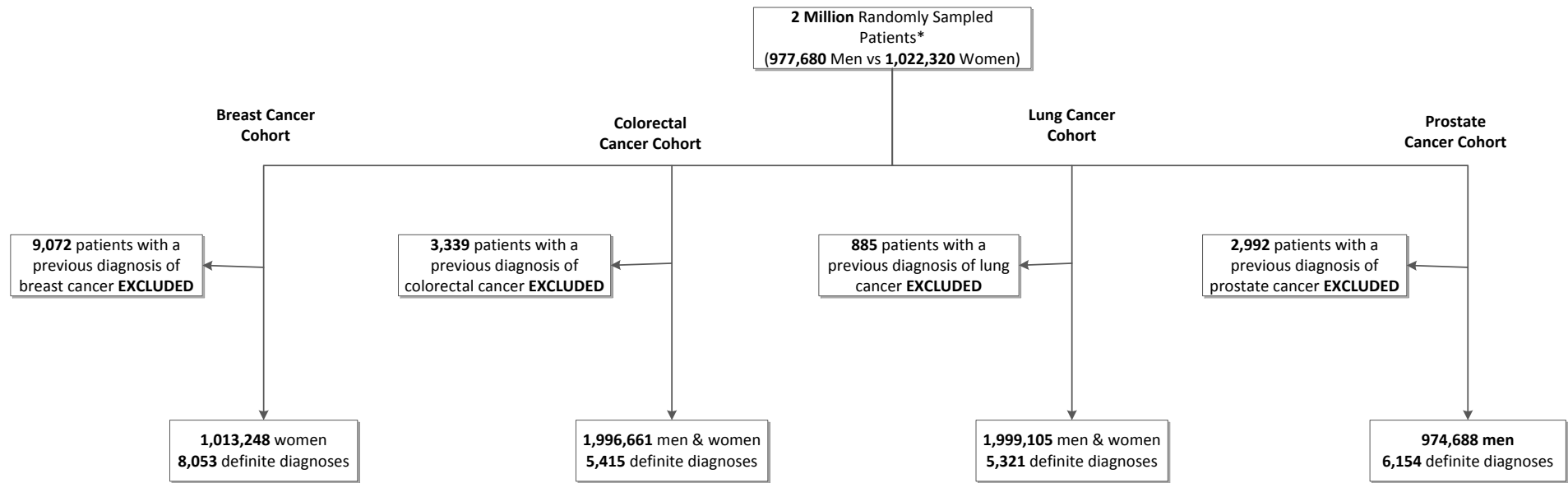
#### 5.3.7.5 Comparisons of incidence rates

Several comparisons of incidence rates were made: first, directly standardised CPRD age-standardised incidence rates based on the European Standard population (pre-2013) were calculated (2000-2010) and compared to directly standardised rates published by the ONS.<sup>182</sup> Overall ONS rates (2000-2010) were not made available to the public. Therefore, a static population over the whole study period (2000-2010) was assumed; derived ONS incidence rates were estimated for each age category by averaging denominator follow-up time. Second, crude age and sex specific rates were compared from the CPRD were compared to reported crude ONS rates. Last, estimated age-standardised incidence rates that incorporated linked data from the NCDR and ONS death registry to the CPRD were compared to national rates provided by the ONS.

**Table 5.1: Distribution of age (at cohort entry) and sex for the random sample of 2 million patients from the CPRD**

Age	Male		Female	
	N	%	N	%
<b>0-4</b>	144 426	(14.8)	136 103	(13.3)
<b>5-9</b>	56 725	(5.8)	52 459	(5.1)
<b>10-14</b>	49 768	(5.1)	44 929	(4.4)
<b>15-19</b>	49 408	(5.1)	51 797	(5.1)
<b>20-24</b>	72 664	(7.4)	94 079	(9.2)
<b>25-29</b>	88 871	(9.1)	101 867	(10.0)
<b>30-34</b>	91 902	(9.4)	91 202	(8.9)
<b>35-39</b>	81 927	(8.4)	75 800	(7.4)
<b>40-44</b>	69 408	(7.1)	63 199	(6.2)
<b>45-49</b>	58 596	(6.0)	54 594	(5.3)
<b>50-54</b>	51 327	(5.2)	50 170	(4.9)
<b>55-59</b>	43 027	(4.4)	43 308	(4.2)
<b>60-64</b>	36 116	(3.7)	38 547	(3.8)
<b>65-69</b>	29 205	(3.0)	33 005	(3.2)
<b>70-74</b>	22 113	(2.3)	28 212	(2.8)
<b>75-79</b>	15 882	(1.6)	24 579	(2.4)
<b>80-84</b>	9 268	(0.9)	18 175	(1.8)
<b>85+</b>	7 047	(0.7)	20 295	(2.0)
<b>Total</b>	<b>977 680</b>	<b>(100.0)</b>	<b>1 022 320</b>	<b>(100.0)</b>

**Figure 5.1: Flow diagram of cancer cohorts - inclusion and exclusion**



\*Sampled from a pool of acceptable patients with at least 6 months up-to-standard (UTS) during the study period (2000-2010)

## 5.4 Results

### 5.4.1 Cohort

The initial study population consisted of 2 million eligible patients (1,022,320 female and 977,680 male) randomly sampled from all eligible CPRD patients (July 2012 version). **Table 5.1** shows the age (at cohort entry) and sex distribution of the 2 million randomly sampled patients. In each 5-year age category, patients were similarly distributed by gender. A significantly larger proportion of patients aged 0-4 years (14.8% male, and 13.3% female), relative to other ages, entered the cohort (latest of 6 months UTS or January 1, 2000).

Four cohorts were formed, one for each individual outcome studied (**Figure 5.1**). Of the 2 million patients, 9,072 patients were excluded due to a previous diagnosis of breast cancer; 3,339 due to a previous diagnosis of colorectal cancer; 885 due to a previous diagnosis of lung cancer; and 2,992 due to a previous diagnosis of prostate cancer.

### 5.4.2 Potential cases and evidence of diagnosis

**Figure 5.2** shows the number of potential cases identified based on the certainty of diagnosis and evidence of diagnosis found in their GP records. For all cancer types, the majority of potential cases were identified using definite codes: 8053/9942 (81%) breast cancer; 5,415/7,103 (76%) colorectal cancer; 5,321/7,180 (74%) lung cancer; and 6,154/7,614 (81%) prostate cancer. Potential cases from “probable” and “possible” diagnostic groups required evidence of diagnosis to confirm case status based on the diagnostic algorithm described in **Chapter 4; Figure 4.1**. The exact forms of evidence of diagnosis varied by cancer type, surgery was most prominent among definite breast and colorectal cancer cases 54.6% and 50.3%,

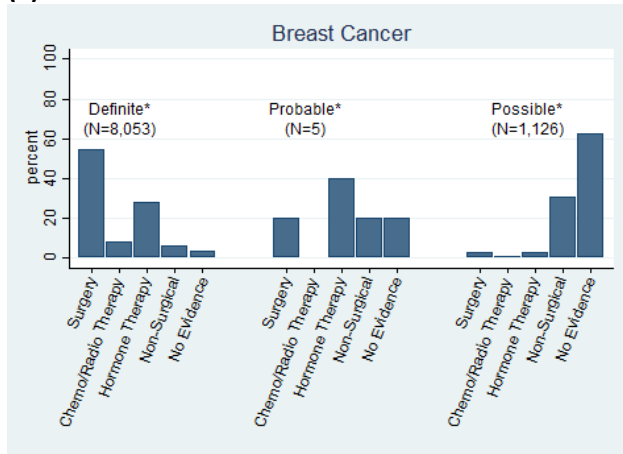
respectively. In contrast, hormonal therapy (46.9%) was more common among prostate cancer cases, while non-surgical evidence (43.2%) such as oncology clinic visits was highest for lung cancer (**Figure 5.2**).

Although cases from both “probable” and “possible” diagnostic groups represented about a quarter of all potential cases (proportion range: 19-24%), overall proportions with evidence of diagnosis (within a 1-year window) to confirm case status was low: breast cancer (42%); colorectal cancer (9%); lung cancer (13%); and prostate cancer (31%) (**Figure 5.2**).

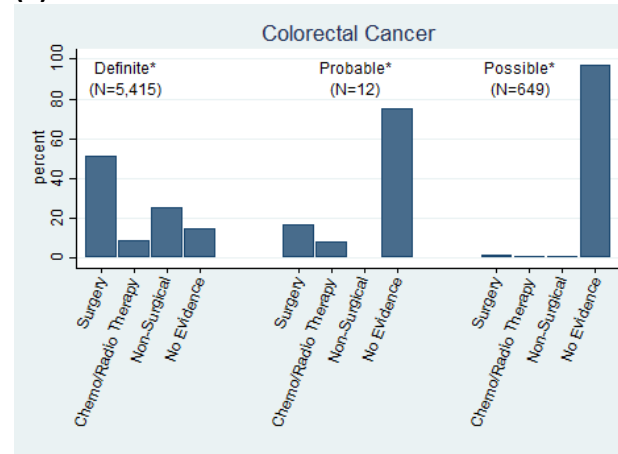


Figure 5.2: Number of potential cases by diagnostic group and corresponding proportion of evidence of diagnosis

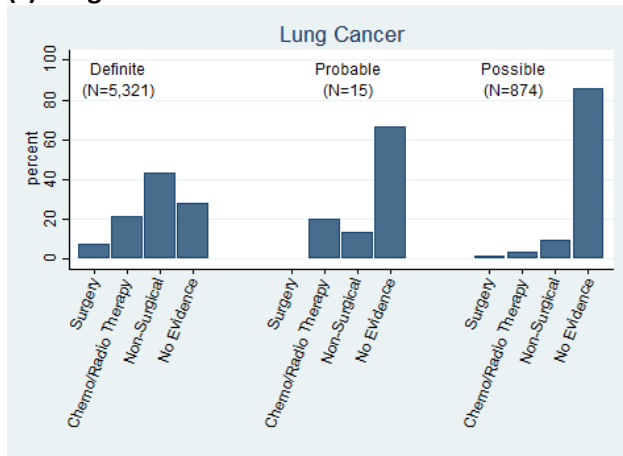
(a) Breast Cancer



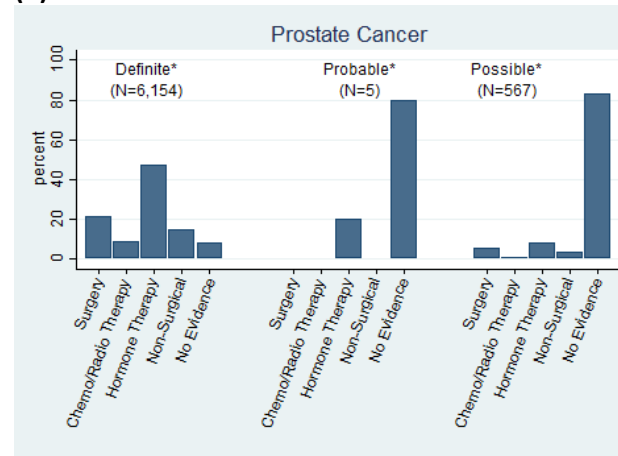
(b) Colorectal Cancer



(c) Lung Cancer



(d) Prostate Cancer



\*Diagnostic groups - **Definite**: patients with a recorded malignant diagnosis; **Probable**: patients with a borderline diagnosis code and evidence of diagnosis; **Possible**: patients with a recorded suspected or general diagnosis code and evidence of diagnosis (Figure 4.1, Chapter 4)

**Table 5.2: Comparison of CPRD age-standardised incidence rates (2000-2010) compared to published ONS age-standardised rates, by cancer type and case definition**

Case definition	CPRD estimated age-standardised incidence rates (ASIR)				ONS estimated incidence rates <sup>c</sup>
	Number of cases	Patient-years	ASIR per 100 000 PY	95% CI	ASIR per 100 000 PY
<b>Breast (C50) Standard definition<sup>a</sup></b>	8,053	5,744,585	111.7	110.5, 113.0	122.9
<b>Broad case definition<sup>b</sup></b>	8,459	5,742,335	119.1	117.8, 120.4	
<b>Colorectal (C18, C19, C20) Standard definition<sup>a</sup></b>					56.4 35.7
Male	3,022	5,712,070	42.9	42.1, 43.6	
Female	2,393	5,814,819	27.8	27.3, 28.4	
<b>Broad case definition<sup>b</sup></b>					
Male	3,031	5,711,739	43.1	42.3, 43.8	
Female	2,400	5,814,459	28.5	27.9, 29.0	
<b>Lung (C34) Standard definition<sup>a</sup></b>					61.1 36.4
Male	3,010	5,724,253	39.6	38.9, 40.3	
Female	2,311	5,826,312	24.4	23.9, 24.9	
<b>Broad case definition<sup>b</sup></b>					
Male	3,091	5,724,093	41.9	41.1, 42.6	
Female	2,362	5,826,182	26.7	26.1, 27.2	
<b>Prostate (C61) Standard definition<sup>a</sup></b>	6,154	5,694,732	84.7	83.6, 85.8	99.7
<b>Broad case definition<sup>b</sup></b>	6,729	5,692,703	86.2	85.1, 87.3	

<sup>a</sup> Standard case definition: includes “definite” cases with a Read code indicating a malignant neoplasm of the cancer of interest

<sup>b</sup> Broad case definition: includes cases from all diagnostic groups – “definite”, “probable”, and “possible” (**Chapter 4, Section 4.2.2**)

<sup>c</sup> Office for National Statistics (ONS)

### **5.4.3 Comparison of incidence rates using CPRD-only case definitions vs ONS published rates**

#### **5.4.3.1 Age-standardised incidence rates**

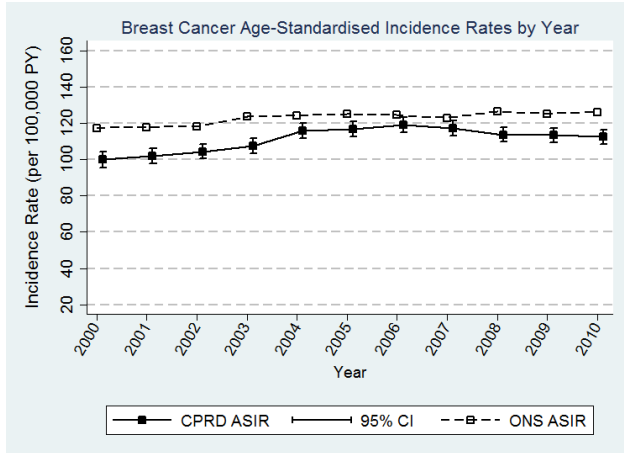
**Table 5.2** shows CPRD age-standardised cancer incidence rates compared to ONS reported rates between 2000-2010. Lower CPRD age-standardised incidence rates (standard case definition) were observed for each of the four cancers under investigation in comparison to ONS reported rates: breast cancer, 111.7 compared with 122.9 per 100k PY (9.1% difference); colorectal cancer, 42.9 compared with 56.4 per 100k PY for men (23.9% difference), and 27.8 compared with 35.7 per 100k for women (22.1% difference); lung cancer, 39.6 compared with 61.1 per 100k PY for men (35.2% difference), and 24.4 compared with 36.4 per 100k PY for women (33.0% difference); prostate cancer, 84.7 compared with 99.7 per 100k PY (15.0% difference). In addition, upper limits of all CPRD age-standardised incidence rate confidence intervals were also lower compared to ONS reported rates.

#### **5.4.3.2 Case definitions**

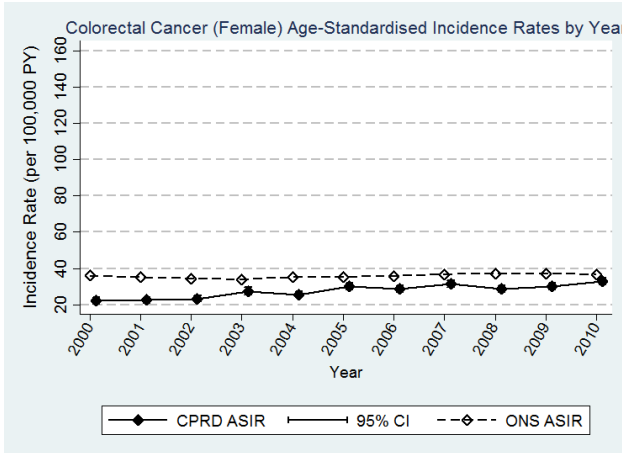
Comparison of the “broad” case definition to that of the “standard” definition resulted in a marginal increase in CPRD incidence rates for all cancer types under study (**Table 5.2**), but yielded mixed results when compared to ONS reported age-standardised rates. Percentage increase for age-standardised incidence rates from standard to broad case definition were slightly higher for breast (6.6%) and lung cancer (Female: 9.4%) were calculated. In contrast, “broad” case definitions yielded slightly lower percentage increase in incidence rates for colorectal (Male: 0.4%; Female: 2.5%) and lung (Male: 5.8%), and prostate cancer (1.7%). All broad case definition age-standardised incidence rates remained lower compared to ONS reported rates.

**Figure 5.3: Comparison of CPRD and ONS age-standardised incidence rates by calendar year**

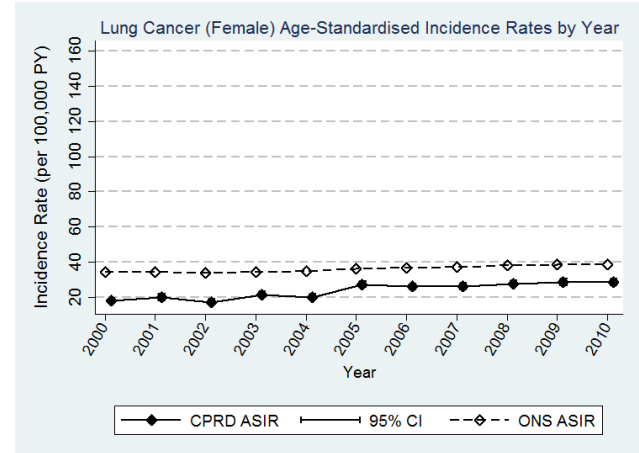
**(a) Breast Cancer**



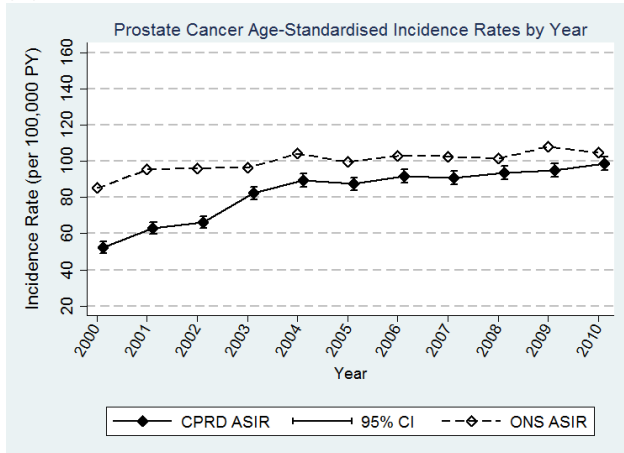
**(c) Colorectal Cancer (Female)**



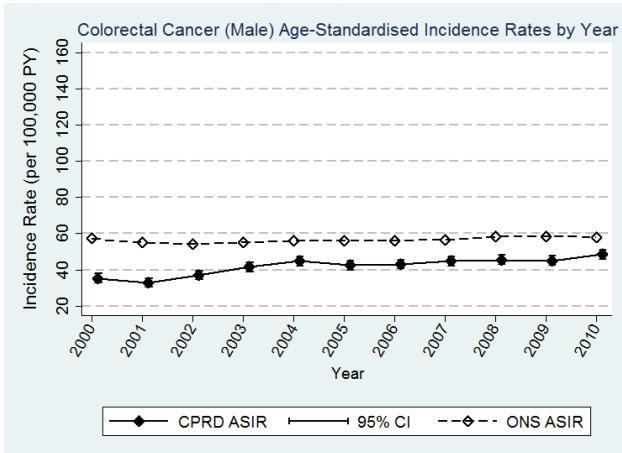
**(e) Lung Cancer (Female)**



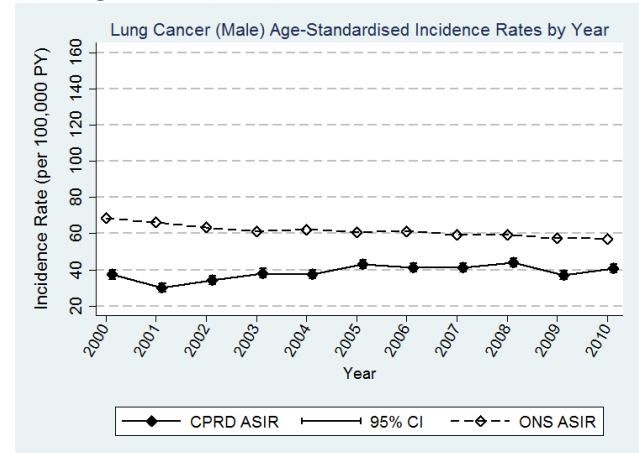
**(b) Prostate Cancer**



**(d) Colorectal Cancer (Male)**



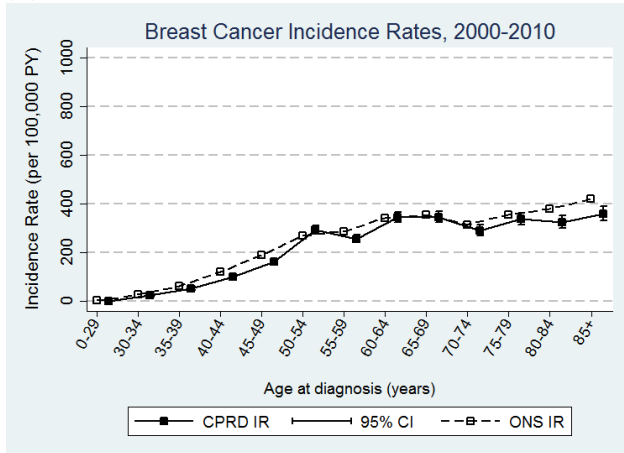
**(f) Lung Cancer (Male)**



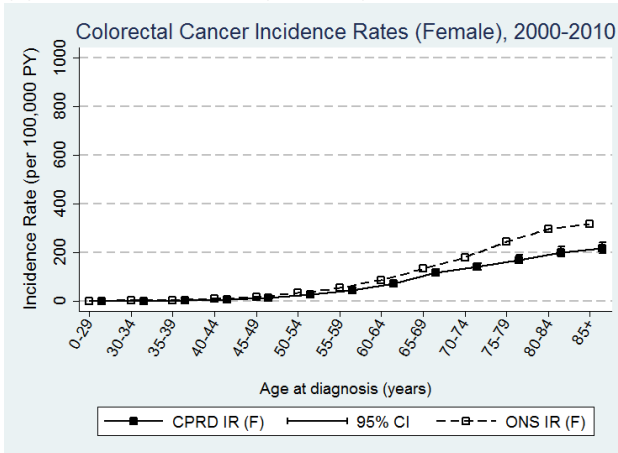
CPRD: Clinical Practice Research Datalink; CI: Confidence Interval; ASIR: Age-Standardised Incidence Rate; ONS: UK Office for National Statistics

**Figure 5.4: Comparison of primary care and ONS reported crude age-specific incidence rates (over the whole study period) by age at diagnosis**

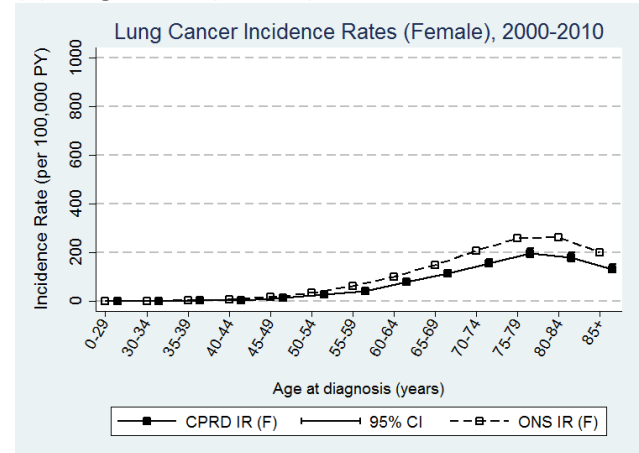
**(a) Breast Cancer**



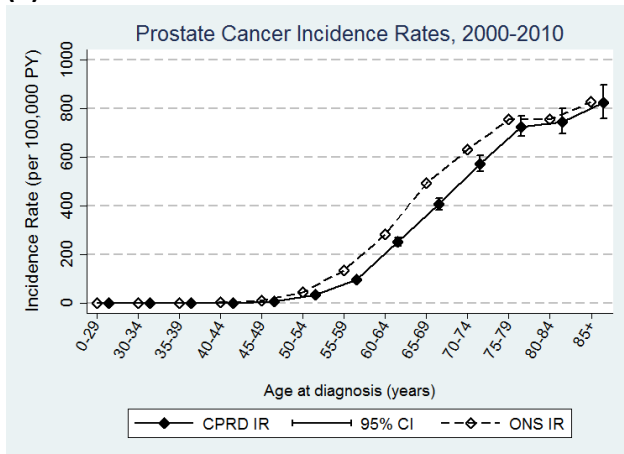
**(c) Colorectal Cancer (Female)**



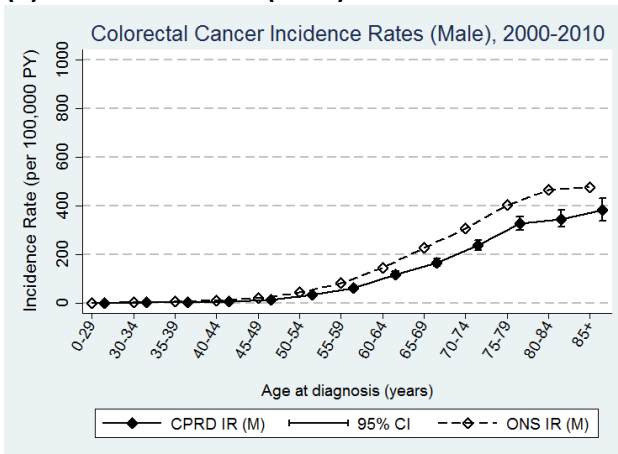
**(e) Lung Cancer (Female)**



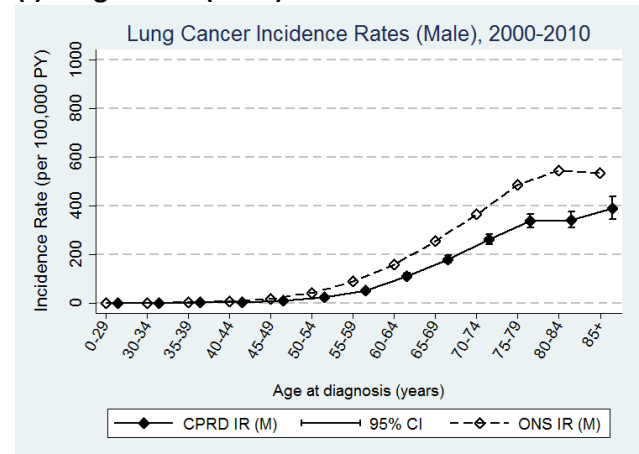
**(b) Prostate Cancer**



**(d) Colorectal Cancer (Male)**



**(f) Lung Cancer (Male)**



CPRD: Clinical Practice Research Datalink; CI: Confidence Interval; IR: Crude Incidence Rate; ONS: UK Office for National Statistics

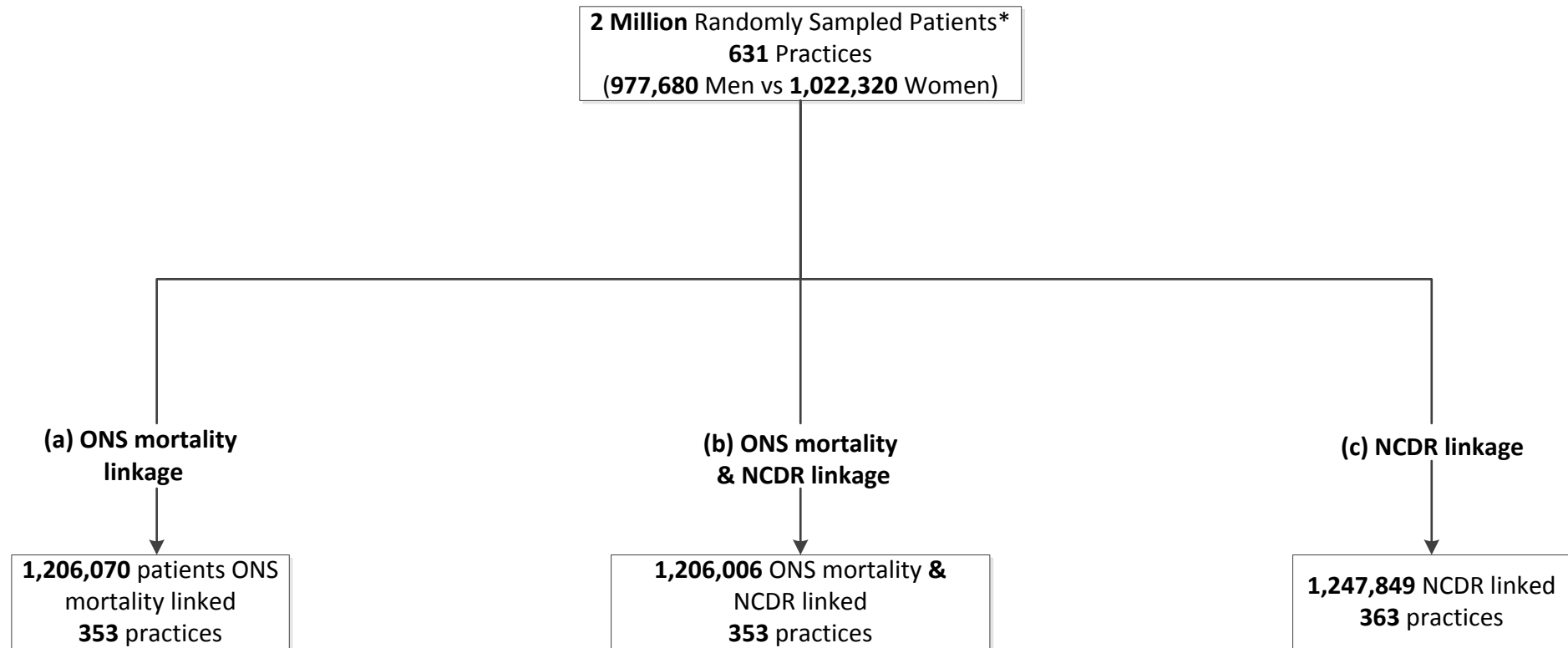
#### **5.4.3.3 Comparison of CPRD and ONS age-standardised incidence rates over time**

**Figure 5.3** shows a comparison of CPRD estimated age-standardised incidence rates for breast, prostate, colorectal, and lung cancer across all ages compared to ONS reported age-standardised incidence rates for each year during the study period (2000-2010). Levels of disparity between CPRD and ONS rates varied by cancer type over time. Differences were particularly pronounced during earlier years of the study period. CPRD and ONS reported age-standardised incidence rates for breast and prostate cancer converged post-2002 and 2003, respectively. For women, colorectal cancer incidence rates converged from 2005 onwards, while a constant difference remained throughout the study period for colorectal cancer in men. This disparity was particularly prominent for lung cancer among both men and women - confidence intervals of estimated rates did not crossover with ONS reported rates over the entire study period.

#### **5.4.3.4 Comparison of CPRD and ONS incidence rates by age at diagnosis**

CPRD crude cancer incidence rates stratified by age were estimated over the study period (**Figure 5.4**). Differences between the CPRD and ONS were found for all cancers as age increased, however, the magnitude of disparity varied by cancer type. Breast cancer incidence rates estimated from the CPRD were similar to ONS reported rates across ages <80 years, although a slight disparity was observed for ages >80 years. Colorectal and lung cancer displayed larger differences compared to breast and prostate cancer, particularly for patients  $\geq 70$  years of age. CPRD prostate cancer incidence rates were lower for men aged 60-74 years, but were relatively similar to those reported by the ONS for men aged 80+ years.

Figure 5.5: Flow diagram of eligible CPRD (primary care) patients linked to the cancer registry (NCDR) and death registry (ONS mortality)



\*Sampled from a pool of acceptable patients with at least 6 months UTS during the study period (2000-2010)

UK Office for National Statistics (ONS)

National Cancer Data Repository (NCDR)

#### **5.4.4 CPRD linkage to the NCDR and ONS mortality**

Of the 2 million CPRD patients, 1,247,849 were eligible for linkage to the cancer registry, 1,206,070 were eligible for linkage to the ONS death registration, and 1,206,006 patients were eligible for linkage in both data sources (**Figure 5.5**).

##### **5.4.4.1 Concordance between CPRD and NCDR**

In total, 5,562, 4,337, 4,838, and 4,717 cases of breast, colorectal, lung, and prostate cancer were identified from all data sources: linkage of the CPRD to the cancer registry and ONS mortality data respectively (**Figure 5.6**). PPVs of CPRD recorded diagnoses against the NCDR was high: breast (89%), colorectal (90%), lung (86%), prostate (88%). In the other direction, sensitivity of CPRD for capturing NCDR recorded diagnosis was lower: breast (89%), colorectal (73%), lung (67%), prostate (81%). Notably, additional linkage to the death registry identified a relatively small number of additional cases for most cancer types not identified in the cancer registry; however, 321 (7%) lung cancer cases were identified in ONS mortality alone. Overall, a low number of patients from the death registry alone were identified, with percentages ranging from 0.4% for prostate cancer to 6.6% for lung cancer.

Cases identified in the cancer registry or death registry but not in the CPRD also varied by cancer type (**Figure 5.6**). For colorectal and lung cancer, 1104/3233 (34%) and 1532/3155 (49%) additional cases were identified in the NCDR and ONS mortality data respectively, while lower proportions were observed for breast 599/4963 (13%) and prostate cancer 827/3890 (21%).

Greater than half of the cases that had a recorded diagnosis in the cancer registry alone had a cancer related diagnosis code in the CPRD (**Table 5.3**). Most of the



cases identified in the NCDR alone were elderly patients, median age at diagnosis ranged from 63-74 years across the four cancer types. Of note, 880 (31%) “NCDR only” lung cancer cases had a recorded death within 90 days of diagnosis, this proportion increased to 40% when extending the time-window to 1 year.

Furthermore, timing of lung cancer diagnosis was near to the date of CPRD recorded end of follow-up for the majority of lung cancer cases identified in the NCDR alone: median 84 days, IQR (26, 836 days) (**Table 5.3**). For all cancer types, few CPRD cases with no matching diagnosis in the NCDR had a cancer diagnosis record of any type in the NCDR (**Table 5.4**). For colorectal (13%) and lung cancer (29%) a substantial proportion of cases died within 1 year of CPRD diagnosis date, this proportion was lower for breast (3%) and prostate cancer (6%).

**Figure 5.6: Agreement of recorded diagnosis between the CPRD, NCDR, and ONS mortality by cancer type**



CPRD: Clinical Practice Research Datalink

**Table 5.3: Cases identified in the cancer registry alone**

	Breast		Colorectal		Lung		Prostate	
	N	(%)	N	(%)	N	(%)	N	(%)
<b>NCDR Only</b>	417		621		301		603	
<b>Any cancer related diagnosis code in CPRD</b>	<b>254</b>	<b>(60.9)</b>	<b>367</b>	<b>(59.1)</b>	<b>184</b>	<b>(61.1)</b>	<b>405</b>	<b>(67.2)</b>
<b><i>Cancer related diagnosis (same site)</i></b>	<b>138</b>	<b>(33.1)</b>	<b>71</b>	<b>(11.4)</b>	<b>35</b>	<b>(11.6)</b>	<b>200</b>	<b>(33.2)</b>
<i>In situ</i>	121	(29.0)	67	(10.8)	8	(2.7)	178	(29.5)
<i>Borderline</i>	1	(0.2)	2	(0.3)	-	-	-	-
<i>Suspected</i>	16	(3.8)	2	(0.3)	27	(9.0)	22	(3.6)
<b><i>Non-specific or Other (not cancer of interest) diagnosis in CPRD</i></b>	<b>116</b>	<b>(27.8)</b>	<b>296</b>	<b>(47.7)</b>	<b>149</b>	<b>(49.5)</b>	<b>205</b>	<b>(34.0)</b>
<i>Malignant related code in the CPRD</i>	0	(0.0)	22	(3.5)	11	(3.7)	28	(4.6)
<i>Non-specific (No site - C80)</i>	55	(13.2)	153	(24.6)	76	(25.2)	104	(17.2)
<i>Other</i>	61	(14.6)	121	(19.5)	62	(20.6)	73	(12.1)
<b>No record of a cancer related code in the CPRD</b>	<b>163</b>	<b>(39.1)</b>	<b>254</b>	<b>(60.9)</b>	<b>117</b>	<b>(28.1)</b>	<b>198</b>	<b>(47.5)</b>
<hr/>								
<b>Death in CPRD</b>								
within 90 days of NCDR diagnosis	13	(3.1)	103	(16.6)	93	(30.9)	27	(4.5)
within 1 year of NCDR diagnosis	22	(5.3)	139	(22.4)	121	(40.2)	53	(8.8)
<b>Time (days) from NCDR diagnosis to end of follow-up: median (IQR):</b>	1211	(533, 2477)	913	(112, 2124)	84	(26, 836)	1231	(503, 2371)
<b>Age at NCDR diagnosis: median (IQR)</b>	63	(54, 76)	74	(64, 82)	73	(64, 80)	71	(64, 78)

**Table 5.4: Cases identified in the CPRD alone (potential CPRD false positive cases)**

	Breast		Colorectal		Lung		Prostate	
	N	(%)	N	(%)	N	(%)	N	(%)
<b>CPRD Only</b>	532		273		345		456	
<b>&gt;1 code for same diagnosis within the CPRD</b>	112	(21.05)	46	(16.85)	39	(11.30)	96	(21.05)
<i>2-4 codes</i>	106	(19.92)	45	(16.48)	39	(11.30)	84	(18.42)
<i>5-9 codes</i>	6	(1.13)	1	(0.37)	0	(0.00)	10	(2.19)
<i>&gt;=10 codes</i>	0	(0.00)	0	(0.00)	0	(0.00)	2	(0.44)
<b>Non-malignant* related code&gt;=1</b>	50	(9.40)	10	(3.66)	24	(6.96)	65	(14.25)
<i>2-4 codes</i>	7	(1.32)	1	(0.37)	5	(1.45)	11	(2.41)
<i>5+ codes</i>	0	(0.00)	0	(0.00)	0	(0.00)	1	(0.22)
<b>Any other cancer related diagnosis in CPRD</b>	38	(7.14)	59	(11.09)	94	(17.67)	61	(11.47)
<b>Death in CPRD</b>								
<i>within 90 days of NCDR diagnosis</i>	9	(1.69)	39	(7.33)	80	(15.04)	11	(2.07)
<i>within 1 year of NCDR diagnosis</i>	17	(3.20)	69	(12.97)	152	(28.57)	34	(6.39)
<b>Time (days) from NCDR diagnosis to start of follow-up, median (IQR)</b>	2680	1318-3962	3078	1740-4403	3018	1668-4515	2872	1546-4345
<b>Age at CPRD diagnosis, median (IQR)</b>	62	(53, 72)	71	(61, 80)	72	(62, 79)	76	(68, 83)
<b>Year of diagnosis</b>	2005	2003-2008	2006	2003-2008	2006	2003-2008	2006	2003-2008

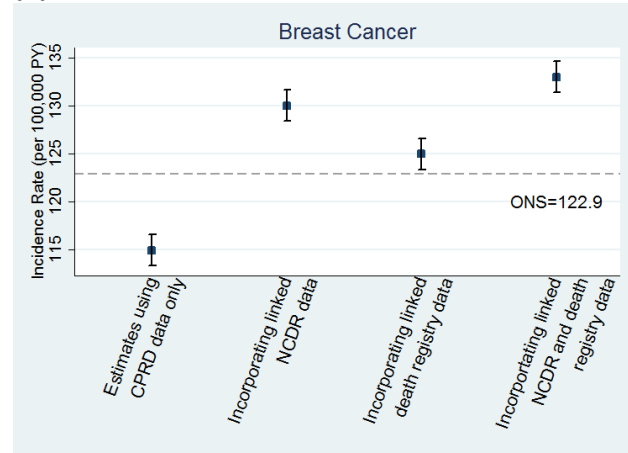
\* Non-malignant: in-situ, borderline, or suspected diagnosis code

#### 5.4.4.2 Overall Incidence rates – linkage to cancer registry and ONS mortality data

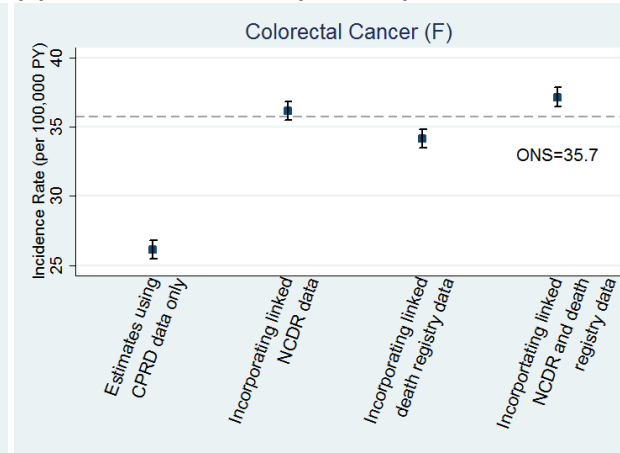
Supplementing CPRD age-standardised incidence rates by the addition of patient level-data from the NCDR and ONS mortality yielded mixed results. In comparison to ONS reported age-standardised rates, supplemented age-standardised incidence rates that included all cases from the CPRD and NCDR were higher for breast (130.0 per 100k PY, CPRD/NCDR vs 122.9 per 100k PY, ONS) and prostate cancer (105.6 per 100k PY, CPRD/NCDR vs 99.7 per 100k PY, ONS). In contrast, incidence rates for colorectal (**male:** 57.6 per 100K PY, CPRD/NCDR vs 56.4 per 100k PY, ONS; **female:** 36.1 per 100K PY, CPRD/NCDR vs 35.7 per 100K PY, ONS) and lung cancer (**male:** 60.7 per 100K PY, CPRD/NCDR vs 61.1 per 100k PY, ONS; **female:** 37.0 per 100K PY, CPRD/NCDR vs 36.4 per 100K PY, ONS) were similar to ONS estimates (**Figure 5.7**). Of note, incidence rates were similar between linkage participating patients and overall CPRD rates from the 2 million sample (**Figure 5.7**). Linkage to ONS mortality data increased incidence rates, although not to the same degree as incorporating NCDR linkage. Incorporating all linked data sources (CPRD, NCDR, ONS Mortality) gave similar estimates to those from linkage analyses linking the CPRD to the NCDR (**Figure 5.7**).

Figure 5.7: Primary age-standardised incidence rates incorporating cancer and death registry linked data by cancer type and linkage

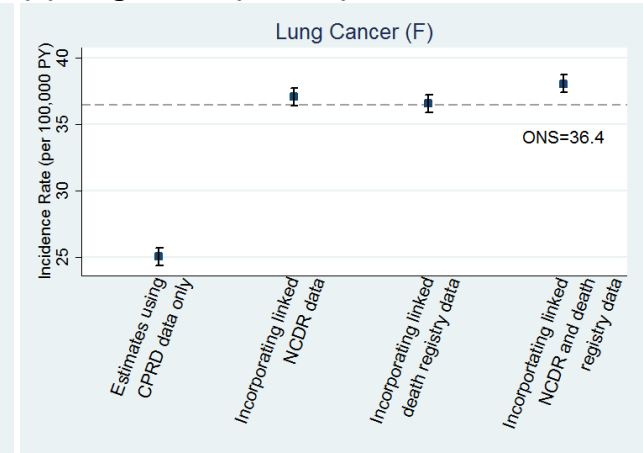
(a) Breast Cancer



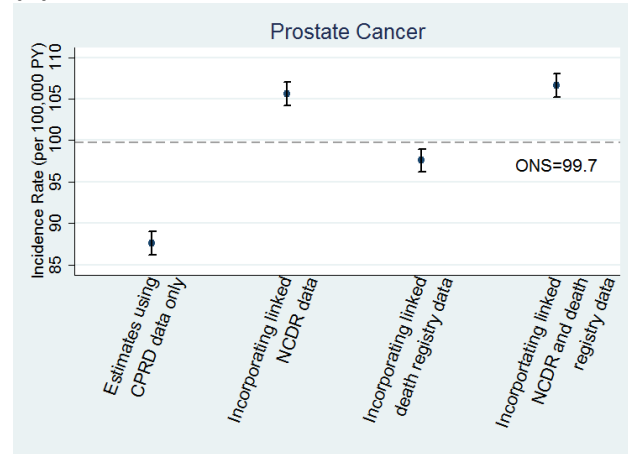
(c) Colorectal Cancer (Female)



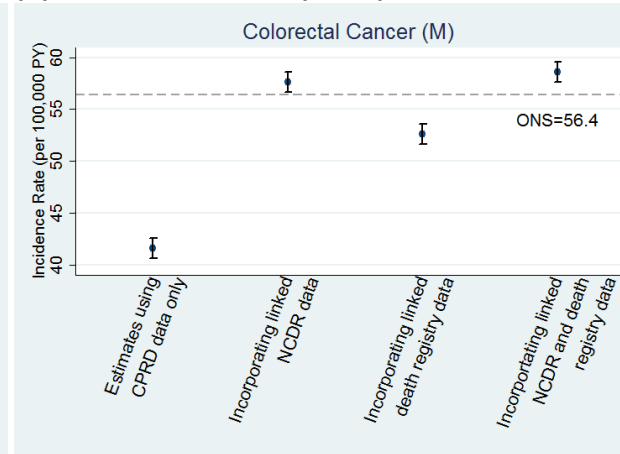
(e) Lung Cancer (Female)



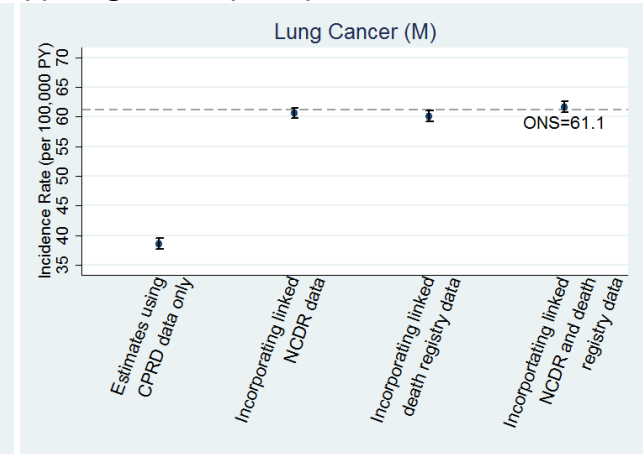
(b) Prostate Cancer



(d) Colorectal Cancer (Male)



(f) Lung Cancer (Male)



CPRD: Clinical Practice Research Datalink; NCDR: National Cancer Data Repository; ONS: Office for National Statistics

## 5.5 Discussion

### 5.5.1 Overview

Incidence rates for four of most common cancer types in the UK from primary care data (CPRD) were compared to national rates published by the ONS. Consistent with previous studies,<sup>187, 188</sup> cancer incidence rates from the CPRD were lower compared to UK national estimates published by the ONS. This was no longer the case when linkage to cancer and death registry data were incorporated: compared to ONS estimates, breast, prostate, colorectal (women), and lung (women) cancer incidence rates were higher than expected, while similar rates were observed for colorectal and lung cancer in men.

### 5.5.2 Comparison of cancer incidence: CPRD vs ONS

Disparities of cancer incidence rates from the CPRD compared to the ONS varied by calendar period, sex, and age. These disparities remained when CPRD and ONS age-standardised incidence rates were compared, which is consistent with the CPRD being representative of the UK population in terms of age and sex.<sup>98</sup> For all four cancer types, the disparity between CPRD and ONS age-standardised incidence rates were particularly larger during the earlier periods of the study (2000-2004); however, estimates from the two data sources converged from 2005 onwards. This trend is consistent with findings from Haynes *et al.*<sup>188</sup>, a possible explanation for this trend could be the implementation of the Quality and Outcomes Framework (QOF) in the UK in 2004 - a programme which provides an incentive to GP surgeries by awarding them achievement points for measures such as: managing chronic diseases, implementing preventative measures, and an overall service of high quality and productivity.<sup>189, 190</sup>

Age at diagnosis contributed to the disparity in incidence rates between the CPRD and ONS, again consistent with findings from past studies.<sup>187, 188, 191</sup> Notably, larger differences were observed for lung and colorectal cancer, particularly for the elderly. Misclassification of case status could have potentially driven this observed relationship: diagnosis of cancer, for some cases, would not have been recorded until death, which may not have been captured in GP records leading to misclassification of some cases as non-cases. Instances where cancer diagnoses are recorded on or near the date of death in CPRD patient records may reflect diseases related to death. Cause of death is not routinely coded in general practice data; however, linkage to the ONS death registry, which collects data from death certificates, does include this information, and can be used in conjunction with the CPRD. Furthermore, this age-related misclassification of case status may be magnified for cancers that are diagnosed at a late stage with low survival rates such as lung or pancreatic cancer.<sup>192</sup> Compared to lung and colorectal cancer, which have higher mortality,<sup>183</sup> disparities by age were not as pronounced for breast and prostate cancers which are typically detected at earlier stages and have higher survival rates.<sup>183</sup>

The disparity observed by cancer type may also have been driven by regional variation in England. Higher incidence rates of breast and prostate cancer have been observed in the South of England (London, South East, and South West), while lower rates were observed in the Northern regions of England (North East & West, and Yorkshire and The Humber). Similarly, lung and colorectal cancer registrations are higher in the Northern regions of England compared to Southern regions.<sup>182</sup>

Variations in regional cancer incidence rates coupled with a higher density of CPRD



GP practices located in the South of England, may have driven the disparity observed between CPRD and ONS published rates.<sup>193</sup>

### 5.5.3 Alternative case definitions

Whether the difference between primary care and registry reported rates would be reduced by using broader case definitions in the CPRD was evaluated. The broad case definition resulted in increased rates for all cancer types, lower rates were still observed for colorectal, lung, prostate cancer; however, similar rates were observed for breast cancer. Recent cancer incidence studies from Haynes *et al.*<sup>188</sup> and Tate *et al.*<sup>191</sup> speculated on whether incidence rates were lower in their studies due to their employment of unambiguous Read codes (definite codes) to define cases. Results from this study and those of Charlton *et al.*<sup>187</sup> suggest that this is not the case. Although additional cases were identified by adopting non-specific cancer codes as part of the case definition, their inclusion resulted in marginal differences to definite CPRD incidence rates for colorectal, lung, and prostate cancer.

Moreover, inclusion of false-positive cases is a likely possibility if a broader case definition is employed. Verification of such cases would be important, for example by request of patient records, free-text, or cancer registry linkage. The last of these verification mediums, namely the cancer registry would be the most pragmatic due to its relative ease to apply to a large sample. Previous studies that have conducted validation studies have done so by requesting hard copies of patient records or accessed free-text, which is a stand-alone facet provided by the CPRD. These methods are resource intensive and have been limited to relatively small case samples as described in the systematic review of cancer outcomes in UK primary care databases (**Chapter 2**). However, results from this study do support the point

that GPs may not always record an event (e.g. prostate cancer diagnosis) by using the most detailed Read code. Findings from this study show that a substantial proportion of cases with a malignant diagnosis in the NCDR have a less specific (borderline, suspected, or malignant diagnosis code with no site specified) recorded diagnosis in primary care records.

In relation to non-specific diagnosis codes, over half of cases identified in the NCDR and not in the CPRD nevertheless had some form of cancer-related diagnosis code within their CPRD records. The proportion of diagnoses recorded in the cancer registry, but not in the CPRD ranged from 10-40% depending on cancer type. However, many of these cases had a record of unspecified malignant neoplasms, in-situ diagnoses or suspected diagnoses.

#### 5.5.4 Linkage to cancer registry and death registrations

In comparison to ONS age-standardised rates, supplemented CPRD age-standardised incidence rates incorporating linked NCDR and death registry data were similar for colorectal and lung cancer, but higher for breast and prostate cancer. It is inevitable that incorporating additional data sources would result in an increased incidence rate unless there was 100% concordance between datasets. Similar rates observed for colorectal/lung cancer and higher rates for breast/prostate cancer could be related to the proportion of potential false-positive (or not registered nationally) cases identified which was higher for breast/prostate (10%) compared to colorectal (6%) and lung (7%) cancer. Cases identified in primary care but not in the cancer registry would necessarily produce higher estimates to reported ONS rates (based on cancer registrations) because of the additional cases identified. However, the magnitude of disparity would depend on the proportion of potential false-negative cases.

Consistent with past studies of CPRD and cancer registry linkage,<sup>80, 194</sup> positive predictive values of CPRD cancer diagnosis were high, ranging from 83% to 89% across all four cancer types. In primary care records, a diagnosis may appear on more than one occasion; for example, a patient may transfer to another practice and the new GP may enter the diagnosis for reference as described by Lewis *et al.*<sup>195</sup> Allowance for this was implemented by excluding any cases that were diagnosed with cancer within 6-months of start of follow-up, and most of the cases identified in the CPRD had been followed for a substantial amount of time before the date of recorded diagnosis in the NCDR.

A small proportion, ranging from 5-10%, of cases were identified in the CPRD but not in either the NCDR or ONS death data. Three possible reasons for cases being identified in the CPRD and not in the NCDR are (i) they are false-positive cases and wrongly recorded as having a cancer diagnosis in the CPRD; or (ii) they could be cases that are missed by the cancer registry; or (iii) were simply never notified of the case from primary care records. 80% of these cases were more likely to be false-positives because only 1 diagnosis code was identified within their CPRD records in comparison to 20% who had >1 definite diagnosis code (Read code mapped to cancer ICD-10) within their CPRD patient records. In addition, the median time from start date to diagnosis date was over 7 years for all cancer types. A short time period from start date to diagnosis date might suggest that the GP entered the diagnosis as a reference point rather than an incident diagnosis.<sup>195</sup> In the other direction, the NCDR may have missed cases that were identified in primary care. Registries receive information on newly diagnosed cases from many sources in the NHS, one of them being general practices. A possibility could be that these primary cases failed ONS validation checks and were not included as registered cases.<sup>182</sup> All coded records in the CPRD typically remain unchanged once recorded, and cases identified would remain as cases within the CPRD.

The sensitivity of the CPRD in capturing nationally registered cancers was lower compared than previous studies reporting on CPRD and cancer registry.<sup>80, 194</sup> Sensitivities ranged from 73% to 89% for breast, colorectal and prostate cancer, while sensitivity for lung cancer was low (66%). The NCDR had a higher number of cases in comparison to cases identified in the CPRD, particularly for cancer types with a lower survival rate such as colorectal and lung cancer.

There are a few possible explanations for lower sensitivities observed in this study compared to previous studies.<sup>80, 194</sup> First, both studies utilised a specific sample of patients from the CPRD: Boggon *et al.*<sup>194</sup> a diabetes cohort, and Dregan *et al.*<sup>80</sup> a cohort of patients with alarm symptoms for cancer. As a result, these groups may be under closer monitoring and therefore would have been more likely to have cancer diagnosed. In contrast, this study randomly sampled from all eligible patients from the CPRD. Second, an earlier version of the linkage set was used in both past studies. Third, Boggon *et al.* also utilised hospital episodes statistics (HES) and free-text data which would increase sensitivity estimates. Last, the CPRD and cancer registry record diagnoses using two different coding dictionaries, the code list used to define cases in both data sources will inevitably influence concordance measures.

The cancer registry collects data on cancer diagnoses from a number of data sources including hospitals, GPs, and coroners, and as such was considered as the gold standard in terms of collection of cancer registrations in the UK. In comparison, the CPRD collects data based on a recording system of administrative physician recording and may not necessarily be complete without supplementation of linkage to external data sources such as the cancer registry. As such, linkage to the cancer registry, death certificates, or hospital data are needed to supplement the existing primary care data. A current disadvantage of the utilisation of linking primary care patient data to a disease specific registry is the trade-off between more comprehensive outcome data and a loss in overall sample size. Nonetheless, one of the main aims of some past epidemiological studies utilising UK primary care

databases on cancer outcomes (**Chapter 2**) was to avoid inclusion of false-positive cases.

### **5.5.5 Cancer Type**

Findings from this study varied by cancer type, however, patterns of consistency emerged among the different cancers examined in this study. Fatality of cancer type, as observed in this study for lung/colorectal cancer compared to breast/prostate cancer, may be a strong predictor of disparity between primary care and ONS reported rates. Similar rates were observed by Haynes *et al.*<sup>95</sup> for cancer types with high survival rates (>90% one year survival for lymphoma, breast, prostate, and melanoma).<sup>188</sup> In comparison, lower primary care incidence rates were observed for pancreatic (21% 1-year survival), lung (32% 1-year survival), colorectal (76% 1-year survival), and ovarian cancer (72% 1-year survival). The exception to these was leukaemia and brain tumours, where similar rates were observed.

### **5.5.6 Limitations**

This study has several limitations. First, findings were limited to the four types of cancer included in this study, which were chosen because they are the four most common cancers diagnosed among men and women in the UK. Second, reference ONS incidence rates over the entire study period (2000-2010) could not be estimated as longitudinal patient data were not available for each calendar year. As such, average overall and age-specific ONS rates across all years were estimated, assuming a static population - averaging reported ONS rates over all years. Year-specific plots (data not shown) showed that there was no apparent change in trend for the individual years. Third, an alternative method to assess data quality could

have been conducted, namely verification of diagnosis through hard-copy requests of GP patient records. Such a validation was not conducted for the cancer diagnosis codes included in this study due to the resource intensive nature of such a process. Moreover, whether GPs that partake in this validation process refer to the same electronic data provided by the CPRD or access extra information is unclear. Optimistic estimates of validity would likely be observed if the former were true. Thus, some misclassification of case status may have been possible which would likely have resulted in an overestimate of incidence rates. However, incidence rates estimated in this study were consistent with past studies. Fourth, not all practices participate in validation or linkage studies, which may limit the generalisability of validity findings. The subgroup of practices included in this study may differ compared to non-participating practices in terms of case file organisation, clarity, or maintenance. Yet, rates were similar between eligible participating practices and overall CPRD rates including all acceptable GP practices. Last, HES and free-text data were not utilised in this study; future studies could incorporate these data sources to assess the impact on incidence rates in comparison to ONS reported rates.

#### **5.5.7 Future Research**

This study was limited to cancers of the breast, colorectum, lung and prostate; future research might investigate the extent to which incidence rates from other cancers vary or are consistent with expected ONS reported rates with the addition of linkages to the NCDR or ONS mortality data. In addition, cancer related epidemiological studies incorporating linkages to external sources could investigate

the impact of age-related biases that might impact study findings if primary care data is used either alone or with the addition of registry data.



## 5.6 Conclusion

Consistent with recent studies, CPRD cancer incidence rates were generally lower compared to ONS reported rates. Incorporating linked data from the cancer registry yielded higher incidence rates for prostate and breast cancer, yet, similar rates, compared to ONS published rates, were observed for cancers of the lung and colorectum.

Three possible permutations of linked data are possible for use in future analysis: (i) incorporating all identified events from all data sources (primary care OR cancer registry); (ii) events restricted to those identified from the cancer registry, as the cancer registry is considered the “gold standard”; and (iii) concordant diagnosis (primary care AND cancer registry) between data sources. Approaches (i) and (ii) capitalise on the gains from incorporating linked data as valid events supplement existing data, whereas approach (ii) limits events to those recorded in primary care. Between approaches (i) and (ii), approach (ii) may be considered the most valid in terms of case ascertainment, as cancer registries apply a stringent validation process to ensure true cases are registered.<sup>182</sup> Moreover, the best approach may depend on the context of the study; in some settings identification of all possible cases is sufficient. In contrast, specificity may be an important property even at the expense of missing some true cases such as pharmacoepidemiological studies **(Chapter 2)**.

In line with previous studies,<sup>196</sup> findings from this study have shown that sole use of primary care databases to identify particular cancer outcomes may be biased and lead to an underestimation of cancer incidence. Primary care data may misclassify case status without external linkage to the cancer registry, particularly among

elderly patients and for cancer types that are captured at a late stage of disease progression with low short-term survival. Failure to incorporate linked data from the cancer and death registry may result in the inclusion of false-negative cases. In the other direction, utilisation of linked data may generate higher cancer incidence either because cases in primary care data may be either false-positive outcomes or are simply not registered nationally. In any case, linkage of primary care data to secondary external data sources, such as the cancer registry and ONS mortality, has proven to be beneficial by allowing exploration of the limitations of primary care data in terms of cancer diagnosis recording.

## **5.7 Summary**

- CPRD (primary care) incidence rates for breast, colorectal, lung, and prostate cancer were lower compared to ONS reported rates based, which were based on cancer registrations.
- Disparities between primary care estimated incidence rates and ONS reported rates varied by cancer type, age at diagnosis, calendar year, and sex.
- High positive predictive value estimates of CPRD recorded diagnoses across all cancer types examined was observed. However, the sensitivity of CPRD for capturing registered cancers was lower.
- CPRD incidence rates incorporating linked cancer registry data yielded similar incidence rates for colorectal (men) and lung cancer (men). However, higher rates were observed for breast and prostate cancer.

## **6 Systematic evaluation of the impact of potential methodological drivers of discrepant results in a pharmacoepidemiological study of statin use and cancer risk**

### **6.1 Introduction**

This chapter describes a series of analyses examining the impact of several methodological aspects of study design in the context of estimating the statin-cancer association within the CPRD.

### **6.2 Objective**

The main objective of this chapter is to measure and compare the individual impact of several potential drivers of discrepant results.

### **6.3 Methods**

#### **6.3.1 Primary methodological outcome measure: assessment of potential drivers of discrepant results**

Several potential drivers of discrepant results were examined in this chapter, which included study bias, alternative outcome definitions, and linkage to the cancer registry and ONS mortality.

Five commonly cited biases that have been noted as potential drivers of discrepant results in previous pharmacoepidemiological studies were examined in this chapter, namely: immortal time,<sup>69</sup> protopathic,<sup>116</sup> prevalent user,<sup>70</sup> healthy user,<sup>118</sup> and time-window bias.<sup>119</sup> In addition, potential factors related to the definition of cases in the CPRD were examined: (i) alternative case definitions and (ii) a comparison of the impact of linking primary care data (CPRD) to the cancer registry to define cancer outcomes.

To address the main objective, several potential drivers of discrepant results were investigated in the CPRD by using a variety of study design methods within the context of estimating the statin-cancer association. For each set of analyses, the relative risk of each cancer of interest was estimated for: **(i)** a design incorporating the potential driving factor (**RR<sub>B</sub>**), and **(ii)** a corresponding “corrected” analysis (**RR<sub>C</sub>**). The difference in log relative risk estimates (**Δβ**; representing change in the un-exponentiated model coefficient of the main treatment effect (statin use) on cancer risk), with corresponding 95% confidence intervals,<sup>197</sup> was the main outcome of interest used to measure the impact of a particular driver within the statin-cancer association, and to give a basic standardised comparison across potential drivers.

The statin-cancer association was selected as a basis for this study due to its importance to public health and additionally the large number of pharmacoepidemiological studies that have presented conflicting findings (**Chapter 3- Systematic Review**). There was assumed to be no causal link between statin use and the risk of cancer based on previous literature and evidence from RCTs,<sup>40, 41, 50, 78, 198</sup> and therefore a valid analysis would yield a confidence interval estimate including 1. In addition, the direction of the risk estimate and consistency between “biased” and corrected analysis was assessed.

A number of analyses and study designs were conducted to examine the impact of potential drivers of discrepant results on the statin-cancer association. Details of these methods are outlined in following sections of this chapter.

### 6.3.2 Outcomes

The primary outcomes of interest include primary incident cancers of the breast, colorectum, lung, and prostate. Specific details of methods related to case identification and definitions are provided in **Chapter 5**.

### 6.3.3 Treatment groups

Overall, three treatment groups were considered for comparison in this chapter: (i) statin users; (ii) statin non-users; and (iii) glaucoma medication users (this was an alternative comparison group used to assess the impact of healthy user bias, **Section 6.3.6.4**). Several treatment definitions were used and are described in the following sections.

#### 6.3.3.1 Statin users

Patients with any statin prescription (British National Formulary (BNF), Chapter 2.12) prior to July 31, 2012 were identified in the CPRD (July 2012 version).

*New statin* users were identified as any patient aged 30-90 years with a first recorded statin prescription during the study period (1995-2012); this time point was defined as the treatment start date. Patients prescribed statins prior to January 1, 1995 were excluded. Furthermore, *new statin* users were required to have at least 12 months UTS registration with their GP before their treatment start date to minimise the likelihood of including prevalent cases of cancer.<sup>199</sup> Statin users with missing date of birth, start date (maximum of either current registration or UTS practice date) or end date (minimum of either death, transfer out, or last collection date) were excluded. Furthermore, statin users with dates that did not agree over follow-up were excluded, for example: start date  $\geq$  end date; date of birth  $\geq$  end date. Patients were also excluded if they had a history of any cancer (malignant

neoplasm, malignant morphology, borderline, in-situ, suspected) prior to the first statin date.

### **6.3.3.2 Potential statin non-users**

CPRD denominator data included all acceptable (e.g. consistent recording of age, sex, registration details and clinical events) patients registered to a GP before July 31, 2012. A pool of eligible (follow-up during the period 1995-2012 and >12 months UTS follow-up) non-users were identified; excluding all patients with a record of a statin prescription. Non-users who went on to become a statin user were included in this pool, their end of follow-up date defined as the day before the first statin prescription. Non-users with dates that did not agree over follow-up were excluded, for example: start date  $\geq$  end date, or date of birth  $\geq$  end date.

### **6.3.3.3 Glaucoma medication users**

Glaucoma medication users were considered as an alternative comparison group to assess the impact of healthy user bias (**Section 6.3.6.4**). Patients with a glaucoma medication prescription (BNF Chapter 11.6) prior to July 31, 2012 were identified in the CPRD (July 2012 version).

New glaucoma medication users were identified as patients aged 30-90 years with no record of glaucoma medication prescription prior to the study period (1995-2012), so that their first prescription (treatment start date) occurred during 1995-2012. Furthermore, new glaucoma medication users were required to have at least 12 months UTS registration with their GP before the equivalent glaucoma medication start date. Glaucoma medication users with dates that did not agree over follow-up were excluded, for example: start date  $\geq$  end date; date of birth  $\geq$  end date.

### **6.3.4 Study design**

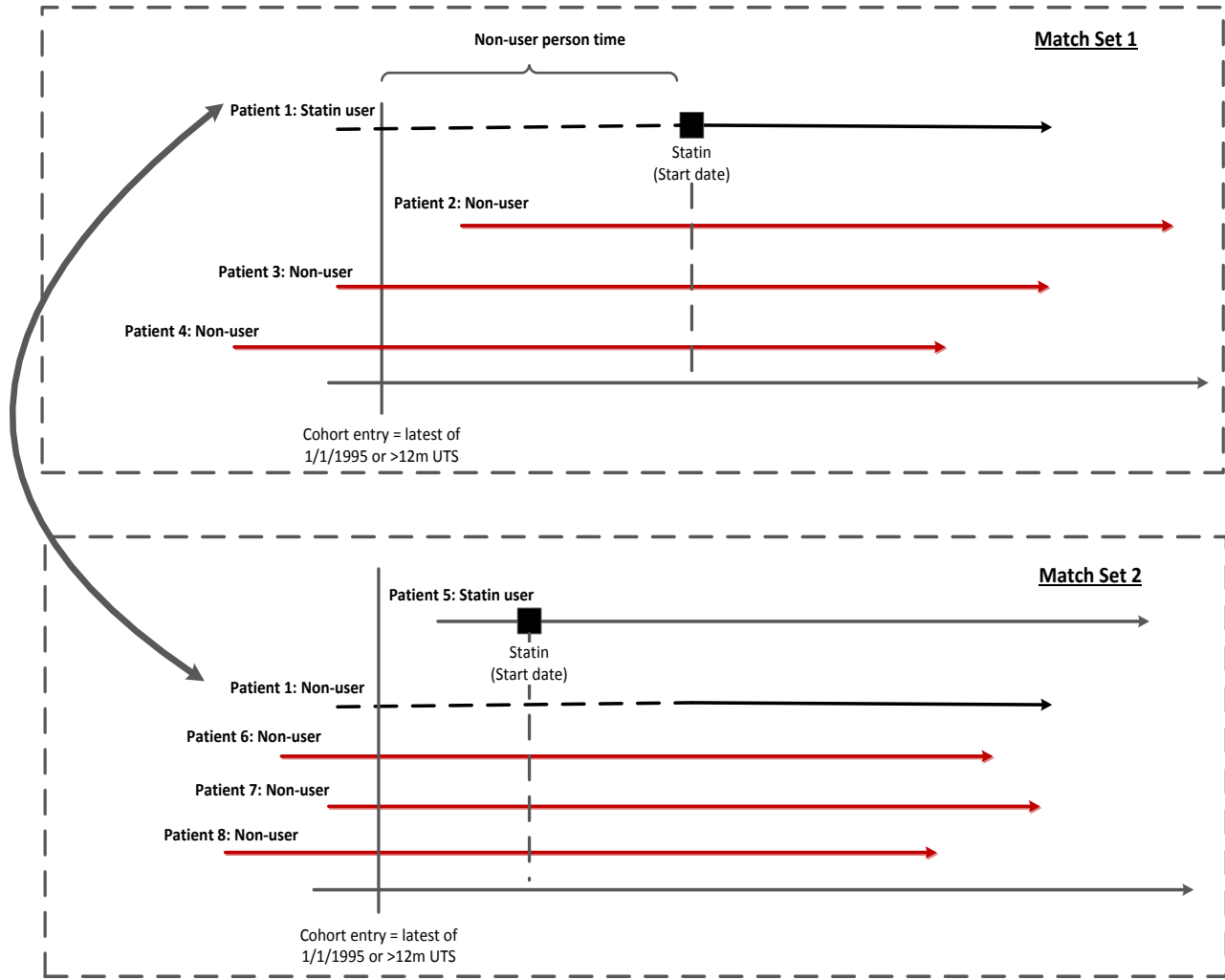
Broad study design methods are described in this section, and specific designs related to the study bias are described in **Section 6.3.6**.

In order to assess the impact of several potential drivers of discrepant results on the risk of cancer among statin users, several study designs were conducted to emulate studies that were conducted in past literature (**Chapter 3 – Systematic review**).

#### **6.3.4.1 Matched cohort of statin and non-users**

From the pool of potential non-users (**Section 6.3.3.2**), all potential matches for each statin user were identified based on the following matching criteria: treatment start date, age ( $\pm 2.5$ ), sex, GP, and >12 months UTS follow-up prior to the matched treatment start date with no history of cancer prior to this date. Sampling of non-users was implemented with replacement, non-users were allowed to be matched to >1 statin user. Once a non-user was selected as a potential control for a user, they would be re-entered into the non-user pool and considered as a potential match for other statin users. Lastly, once a pool of potential non-user matches was selected for each statin user, five non-users (ratio of 5 non-users to 1 statin user) were selected based on closest age difference.

**Figure 6.1: Diagrammatic representation of intention to treat analysis**



**Match Set 1:**  
Patient 1: New statin user

**Match Set 2:**  
Patient 1: Non-user – potential match to Patient 5: statin user

**Intention To Treat (ITT) Analysis**  
In **Match Set 2**, if non-user time from **Patient 1** is matched as a non-user control for **Patient 5**, then the initial treatment group to which **Patient 1** belonged will define treatment status.

Furthermore, since **Patient 1** is a matched non-user for **Patient 5**, **Match set 1** is excluded from further analyses.

Follow-up time for **Patient 1 (Match Set 2)** will start at cohort entry and end at last observed point of contact, ignoring subsequent statin use as depicted in **Match Set 1**

UTS: Up-to-standard



Tim Collier, Krishnan Bhaskaran and Harriet Forbes wrote the original program to implement matching.

Cohort analysis was conducted on an intention to treat (ITT) basis. This design feature is analogous to a randomised design where treatment assignment is based on initial randomisation and not the treatment eventually received.<sup>56</sup> As illustrated in **Figure 6.1**, the initial treatment status of a patient would determine their classification, hence anyone selected as a matched non-user would be excluded as a later statin user.

#### **Observation period: matched cohort**

**Matched cohort (6.3.4.1):** The period of observation for cancer events began at the treatment start date for statin users or corresponding matched treatment start date for non-users up to either the end date or first date of diagnosis for a cancer of interest or censored at a malignant diagnosis of another cancer.

#### **6.3.4.2 Nested case-control study**

##### **6.3.4.2.1 Base cohort**

A cohort of *new statin* users and non-users aged between 30-90 years at cohort entry was identified in the CPRD. Patients were excluded if they had a previous history of cancer prior to cohort entry: latest of January 1, 1995 or start of registration with their GP.

##### **6.3.4.2.2 Case-control: time independent sampling**

Cases of breast, colorectal, lung, and prostate cancer were identified among the cohort of patients. The date corresponding to the first medical record of a definite diagnosis of breast, colorectal, lung or prostate cancer was assigned as the index

date for cases. Controls were defined as patients with no record of cancer prior to their end date (minimum of either death, transfer out date, or last collection date).

#### **6.3.4.2.3 Nested case-control: risk-set sampling**

For the case-control design, patients with no record of a cancer of interest were selected as potential controls. However, the time before diagnosis for an identified case was considered as potential control person time, and therefore cases were eligible to be controls for other cases.

#### **Matching of controls to cases**

Firstly, controls were matched according to index date, age ( $\pm 2.5$  years), sex, GP and >12m UTS follow-up before the index date. Secondly, as controls were matched on index date, all eligible controls with UTS follow-up prior to the matched index date were considered as potential controls for that particular case (risk-set sampling). Controls were allowed to be matched to >1 case (replacement of controls).

#### **Treatment definition and observation period**

A patient was defined as a statin user if they had a recorded prescription 12 months prior to their observed end of follow-up: end date or event of interest.

The period of observation for exposure began 12 months prior to the index date for a case or matched index date for controls, back to the earliest occurrence of follow-up, defined by either the start date or a statin prescription.

## 6.3.5 Statistical analysis

### 6.3.5.1 Statistical models

To model the risk of cancer associated with statin use, three statistical models were used: (i) Cox regression for the matched cohort design (**6.3.4.1**); (ii) conditional logistic regression for the nested case-control design (**6.3.4.2.3**); and (iii) unconditional logistic regression for the case-control design with time-independent sampling (**6.3.4.2.2**). Hazard and odds ratios with corresponding 95% confidence intervals were estimated as appropriate.

In addition, the Cox regression model (**6.3.4.1**) was stratified by “match set” (**Figure 6.1**) enabling equal coefficient estimates across “match set” with individual baseline hazard estimates unique to each stratum.<sup>200</sup>

### 6.3.5.2 Confounders

Age at first statin (cohort) or diagnosis (case-control), sex, calendar year, and general practice were matching variables in both the matched cohort and nested case control design. Calendar year was included in the statistical adjustment of the *new statin* and *new glaucoma medication* cohort. Lifestyle factors included: smoking status (non-smoker, current smoker, ex-smoker, and unknown), body mass index (BMI) (<20, 20-25,>25, unknown), alcohol status (non, ex, current, rare <2u/d, moderate 3-6u/d, excessive >6u/d, unknown), and rate of GP consultation visit rate.

Code lists available from the EHR group at LSHTM were updated and utilised to search CPRD patient records for the following co-morbidities: diabetes, coronary heart disease (CHD), heart failure, hypertension, and hyperlipidaemia. Co-medications included non-steroidal anti-inflammatory drugs (NSAID) or aspirin use

(BNF: 10.1.1). antihypertensive medications: thiazides and diuretics (BNF: 2.2.1), beta-blockers (BNF: 2.4), angiotensin-II receptor blockers and angiotensin-converting enzyme inhibitors (BNF: 2.5.5), calcium channel blockers (BNF: 2.6.2); oral contraceptives (BNF: 7.3.1) and hormone replacement therapy (HRT) (BNF: 6.4.1.1). All co-morbidities and co-medications were formatted as binary variables (No/Yes). All potential confounders were identified within a ( $\pm$ ) 1-year window to the treatment start date (matched date for non-users) for the matched cohort and index date (date of diagnosis) for the nested case-control design. All potential confounders were compared and tested for differences between treatment groups (case status for nested case-control design) by a Chi-squared test for categorical variables and a t-test for continuous variables. Corresponding p-values were two-sided.

### **6.3.5.3 Sensitivity analysis**

#### **6.3.5.3.1 Matching with replacement: weighting of non-users**

Non-statin users were matched to  $\geq 1$  statin user. In the statistical models utilised, all non-users were considered independent, however potential non-users could be included in the analysis on more than one occasion and the independence assumption of the statistical models may be violated.

In order to account for the non-independence, inverse frequency weights were included in all statistical models that incorporated matching with replacement.

Inverse frequency weights were based on the inverse number of times a non-user was matched to different statin users.<sup>201</sup>

#### **6.3.5.3.2 Missing data**

Complete case analysis was conducted in all primary analyses.<sup>202</sup> Patients with missing data (unknown) on smoking status, BMI, or alcohol status were excluded from further analysis. For the complete case analysis, the probability of missingness in BMI, smoking status, and alcohol status was assumed to be independent of the cancer events conditional on observed covariates measured.<sup>202</sup> Sensitivity analyses were conducted to assess the impact of missing BMI, smoking status, and alcohol data: (i) inclusion of a separate missing data category for each of the incomplete variables; (ii) multiple imputation.

Multiple imputation of missing BMI, smoking status, and alcohol status was implemented using chained equations. The imputation model included all potential confounders included in the main model listed in **Section 6.3.5.2** as well as the cancer event of interest; five imputed datasets were created and the results were combined using Rubin's rules.<sup>203, 204</sup>

### **6.3.5.3.3 Censoring at treatment change**

All analyses that were conducted on the matched cohort were conducted on an intention to treat basis. As a sensitivity analysis, the effect of censoring follow-up was conducted. For each patient, censoring of follow-up was implemented if a patient's exposure status changed. For example, statin users were censored if they stopped statin use for a collective period of 6 months. Similarly, a non-user would be censored if they initiated statin use.

## **6.3.6 Impact of potential drivers of discrepant results**

### **6.3.6.1 Immortal time bias: biased and corrected designs**

#### **Description**

Immortal time bias is introduced when a wait period in which an event cannot occur is implemented within the design of a study. For example, an exposure definition requiring 6-months follow-up from therapy initiation before observation for outcome events can commence. This wait period gives a survival advantage to the exposed group until the treatment definition is met, leading to a spurious protective bias on the risk estimate.<sup>69</sup>

### **Study design utilised**

From the matched cohort of *new statin* users and non-users (**6.3.4.1**) treatment status was defined in two ways:

- (i) Requirement of 2 recorded statin prescriptions during follow-up.
- (ii) Requirement of 2 recorded statin prescriptions within the first 6 months of treatment start date, and a minimum of 6-months follow-up duration after the first statin date.

Statin users (and corresponding matched non-users) who did not meet the treatment definition were excluded from further analyses.

### **Potentially biased design**

For both treatment definitions, the start of follow-up began at the first statin prescription during follow-up (and corresponding matched treatment start date for non-users). However, statin users still needed to satisfy the treatment definition which included the period of immortal time where an event could not occur (**Figure 6.2 (i)**).

### **Correction of immortal time bias**

For the corrected design, immortal time was excluded by starting follow-up at the time point where the treatment definition was satisfied: (i) follow-up began at the second consecutive statin prescription in cohort 1; (ii) follow-up began after a minimum of 6-months follow-up from the first recorded statin prescription (**Figure 6.2 (ii)**).

### **6.3.6.2 Protopathic bias: biased and corrected designs**

#### **Description**

Protopathic bias occurs when patients with latent cancer which has not been diagnosed may present with symptoms that lead to a statin being initiated. For example, due to a pre-existing cancer, changes in diet or physical activity may lead to changes in patient lipid profiles, which in turn may lead to a statin being prescribed. As these individuals are subsequently diagnosed with cancer, statin use may mistakenly be associated as the cause of cancer, when in fact the pre-diagnosis cancer symptoms caused the statin initiation.<sup>116</sup>

#### **Study Design utilised**

Protopathic bias was examined by utilising the *new statin* user matched cohort described in **Section 6.3.4.1**.

#### **Potentially biased design**

The potentially biased analysis did not implement a minimum period of exposure, any events occurring early on during initiation of statin use were included in the relative risk estimate. Start of follow-up began at the date of first statin prescription **(Figure 6.3 (i))**.

#### **Correction of protopathic bias**

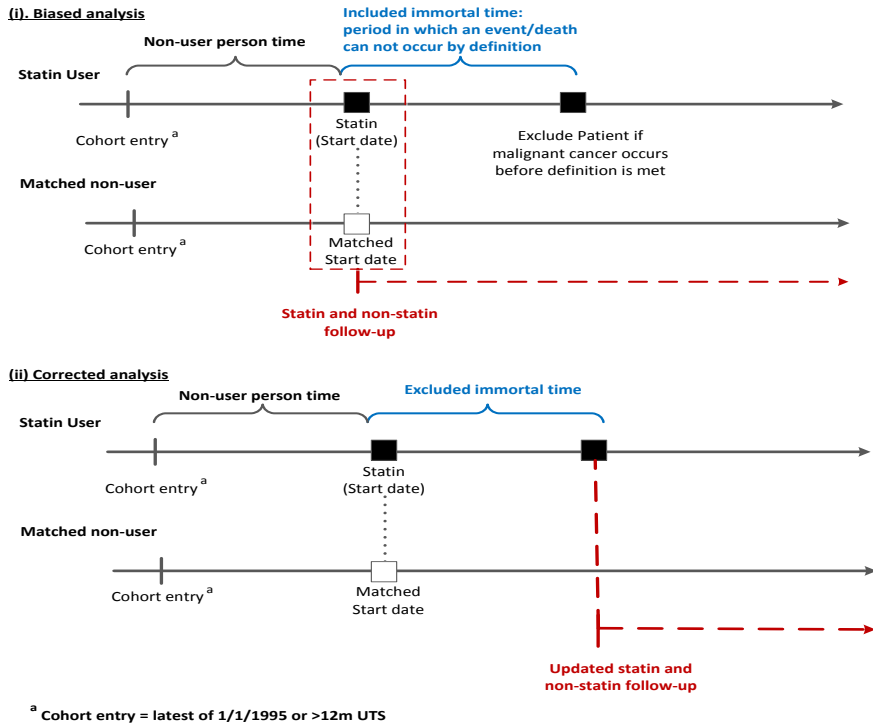
For each cancer type, the relative risk of cancer among statin users from multiple analyses incorporating sequentially increasing minimum periods of exposure (30 day increments) was estimated. The corrected analysis was set at 360-day lag **(Figure 6.1.3 (ii))**.



## Figure 6.2: Immortal time bias: biased and corrected designs

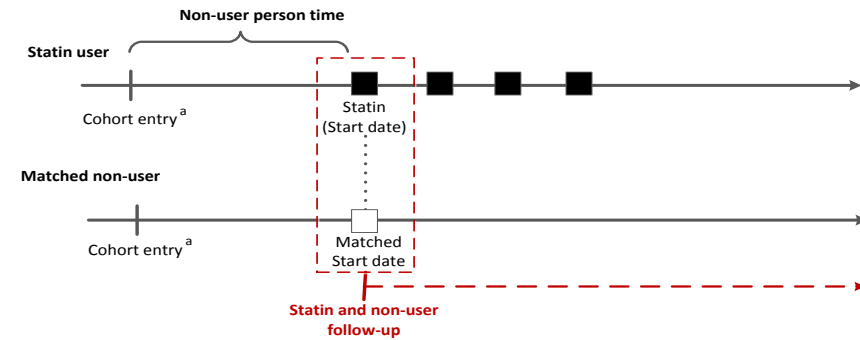
Exposure definition 1: minimum of two recorded statin prescriptions including first statin during follow-up  
 Exposure definition 2: minimum of two recorded statin prescriptions within 6 months of the first statin date (start date) during UTS follow-up

Patients with <2 statin prescriptions (Exposure definition 1) or <2 prescriptions within a 6-month window (Exposure definition 2) from their first statin (start date) will be excluded along with corresponding matched nonusers.

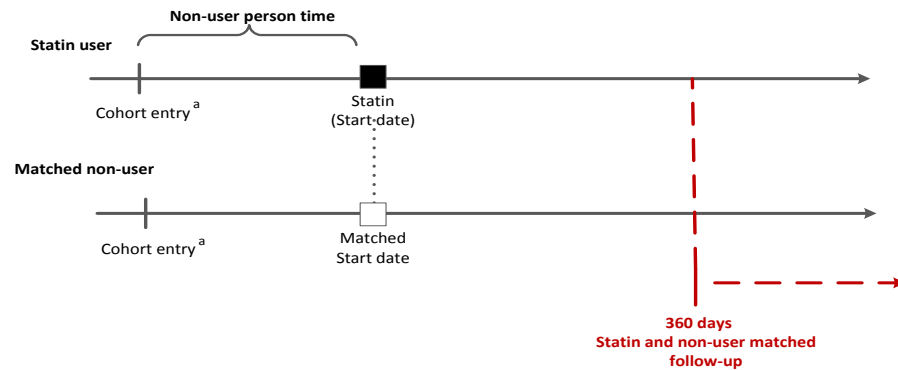


## Figure 6.3: Protopathic bias: biased and corrected designs

**(i) Biased analysis: 0-days lag implemented**



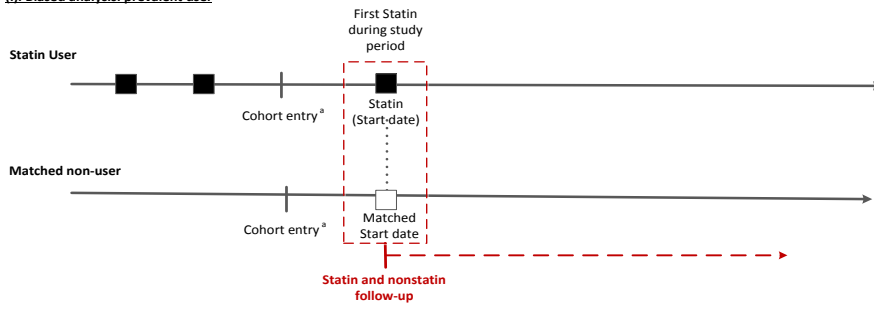
**(ii) Corrected analysis: 360-days lag implemented**



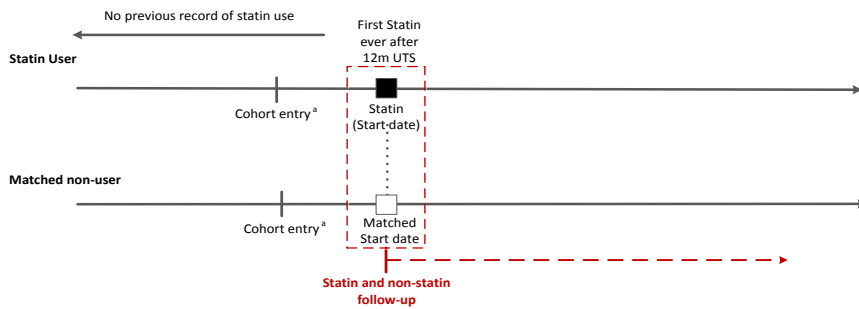
<sup>a</sup> Cohort entry = latest of 1/1/1995 or >12m Up-to-standard follow-up (UTS)

**Figure 6.4: Prevalent User bias: biased and corrected designs**

**(i). Biased analysis: prevalent user**



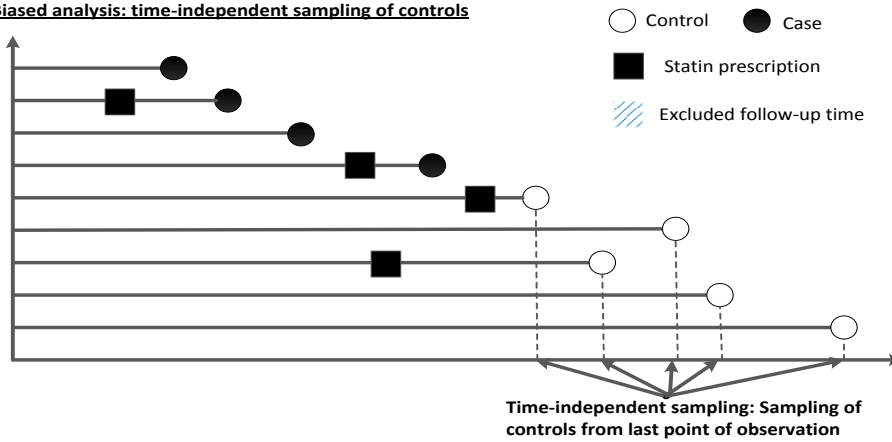
**(ii). Corrected analysis: new user**



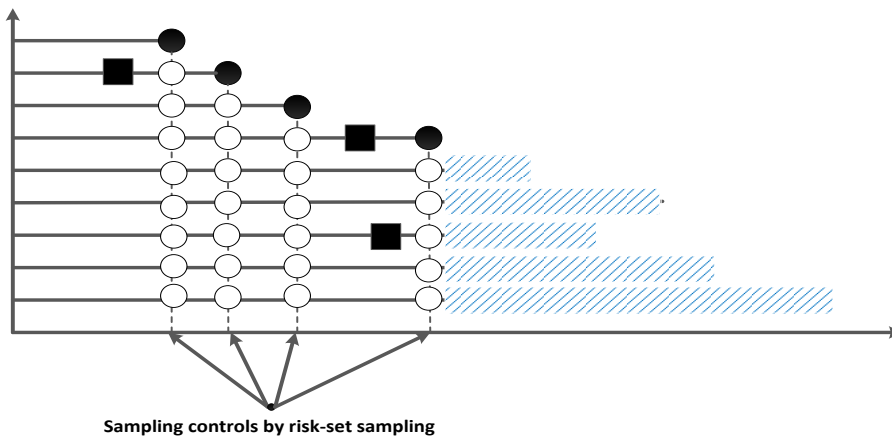
<sup>a</sup> Cohort entry = latest of 1/1/1995 or >12m Up-to-standard follow-up (UTS)

**Figure 6.5: Time-window bias: biased and corrected designs**

**(i) Biased analysis: time-independent sampling of controls**



**(ii) Corrected analysis: risk-set sampling of controls**



### 6.3.6.3 Prevalent user bias: biased and corrected designs

#### Description

Prevalent users of statin therapy may be different from initiators of statins.

Prevalent statin users have remained on therapy for a longer period of time compared to *new statin* users which potentially introduces a number of protective biases. Firstly, depletion of susceptible patients, in which prevalent users would include patients that have persisted with statin therapy, and remained tolerant of any potential side effects from the drug. Secondly, prevalent users of medications are associated with better adherence and outcomes overall compared to new users.<sup>70</sup>

#### Study designs utilised

Two sets of analyses were used to assess the impact of prevalent user bias: (i) a matched cohort of *new statin* users and non-users (**Section 6.3.4.1**); and (ii) a matched cohort of statin users (incident and prevalent) and non-users.

The cohort of *any* statin users (incident and prevalent) matched to non-users was formed to assess the impact of prevalent user bias. The main differences between the *any statin* cohort and the *new statin* cohort included the following:

1. Requirement for >12 months UTS follow-up
2. Inclusion of prevalent statin users: defined as any statin use prior to the start of cohort entry (January 1, 1995)

The period of observation for cancer events began at the treatment start date for statin users or corresponding matched treatment start date for non-users up to either the end date or first date of diagnosis for a cancer of interest or censored at a malignant diagnosis of another cancer.

### **Potentially biased design**

The relative risk of cancer among patients with any record of statin use (incident and prevalent) was estimated.

### **Correction of Prevalent user bias**

To correct prevalent user bias, an analysis of *new statin* users was conducted and compared to findings from the prevalent user analysis (**Figure 6.4 (ii)**).

### **6.3.6.4 Healthy user bias: corrected and biased designs**

#### **Description**

Statins are a widely used class of drug prescribed to lower cholesterol levels, and aid in the management and prevention of stroke. Similarly, glaucoma medications are used to prevent progression of glaucoma. Compared to non-users, patients prescribed preventative therapy may be more likely to have better health-seeking behaviour, such as exercise, healthier diet, and may adhere to health services directed at preventing related diseases, which may influence study findings.<sup>205</sup>

#### **Study design**

In order to assess the impact of healthy user bias two cohorts were compared (i) a matched cohort of *new statin* users and non-users (**Section 6.3.4.1**); and (ii) an unmatched cohort of *new statin* users compared to *new glaucoma* medication users (**Section 6.3.3.3**). The two treatment groups were joined; *new statin* users with a prior record of glaucoma medication use were excluded and vice-versa.

Observation for events began at the treatment start date for *new statin* users and equivalent start date for new glaucoma medication users up to either the end date or first date of diagnosis for a cancer of interest or censored at a malignant diagnosis of another cancer.

### **Potentially biased design**

*New statin* users compared to non-users as described in **Section 6.3.4.1**.

### **Correction of Healthy user bias**

To assess the effect of healthy user bias the corrected analysis consisted of *new statin* users compared to initiators of another preventative medication, namely: glaucoma medications. New users of both medications were defined as patients with a first ever recorded prescription of the drug of interest after cohort entry. Additionally, this was the time-point at which follow-up commenced for both cohorts.

### **6.3.6.5 Time-window bias (nested case-control): corrected and biased designs**

#### **Description**

The time-window used to assess exposure status between cases and controls may not necessarily be fairly distributed if a time-independent sampling strategy of controls is utilised. Cases in general may have a shorter period of observation compared to their counterpart controls. Compared to controls, this would like lead to a shorter observation period to classify treatment status for cases, which would likely lead to an overrepresentation of both exposed controls and unexposed cases leading to a downward bias.<sup>119</sup>

#### **Study designs**

A case-control design which sampled controls independently of time (**Section 6.3.4.2.2**) was used to implement the biased analysis and a nested case-control design (risk-set sampling of controls) was utilised for the corrected analysis (**Section 6.3.4.2.3**).

### **Potentially biased design**

From the case-control design (**Section 6.3.4.2.2**), controls were sampled from their last point of contact in the CPRD independently of time (end of follow-up; end of CPRD follow-up; or death from any other cause). Exposure history was also ascertained up to this time point.

### **Correction of Exposure Time-Window Bias**

Risk-set sampling was implemented to sample and select controls in the corrected design. This allowed a fair observation period between cases and controls to classify treatment status. Controls were sampled from those that were eligible and under follow-up on the day the case was diagnosed (**Figure 6.5 (ii)**).

### **6.3.6.6 Alternative outcome definitions**

#### **Description**

As shown in the systematic review conducted in **Chapter 2**, there has been a divide between research groups that have conducted cancer outcome studies in UK primary care databases. Findings from the systematic review in **Chapter 2** show that the majority of studies have implemented case definitions requiring only a malignant diagnosis code for the cancer of interest. In contrast, other studies have implemented case definitions that have been based on a broader set of diagnosis codes including malignant and non-malignant codes with the requirement of evidence of diagnosis such as cancer related surgery or chemotherapy to confirm case status. Whether alternative case definitions that include non-malignant diagnoses may impact study findings is unknown.

### **Study design**

For both case definitions, the matched cohort of *new statin* users and non-users

**(Section 6.3.4.1)** was used to examine the impact of alternative case definitions on the relative risk of cancer.

### **Standard case definition**

All previous analyses described in this chapter have been based on a “standard” case definition **(Chapter 4, Section 4.2.3)** which required at least one definite malignant diagnosis code during follow-up and no previous history of cancer. Here, the impact of using a broader definition was investigated.

### **Broad case definition**

The broad case definition included patients from three diagnostic groups (i) cases with a definite malignant diagnosis of the cancer of interest; (ii) probable and (iii) possible cases. Probable and possible cases required supporting evidence of diagnosis such as cancer related surgery or chemotherapy to confirm case status **(Chapter 4, Section 4.2.2)**. Early detection of cancers may be differential between treatment groups particularly for patients prescribed statins (disease prevention drug). As demonstrated in **Chapter 5**, some cases identified in the cancer registry alone had a related code entered in their CPRD records i.e. carcinoma in-situ, suspected diagnosis, or a malignant diagnosis (site unspecified). Reasons for these recordings are unclear, whether these diagnoses were entered retrospectively by the GP (prevalent diagnosis) or were a detailed account of disease progression.

### 6.3.6.7 Linkage to the cancer and death registry

#### Description

The impact of incorporating cancer outcomes from cancer registry data within a pharmacoepidemiological setting is unknown. Linkage of primary care data (CPRD) to the cancer registry provides several challenges. Firstly, not all general practices in the CPRD consented to linkage studies, which may imply a difference between analyses utilising all of the CPRD, compared to a subset of the CPRD which was eligible or consented to linkage research. Secondly, the extent of concordance may also affect results: Boggon *et al.*<sup>194</sup> and Dregan *et al.*<sup>80</sup> have shown that all cancers are not equal in terms of concordance. Cancers with high mortality rates such as lung and pancreatic cancer have lower levels of concordance.

#### Study design

Three linkage analyses were conducted: (i) cancer outcomes identified from the CPRD alone; (ii) incorporation of all outcomes from both the CPRD **OR** NCDR; and (iii) restricting cancer events to concordant diagnosis only (CPRD **AND** NCDR). For all analyses, the matched cohort of *new statin* users and non-users (**Section 6.3.4.1**) was used to examine the impact of linking primary care data (CPRD) to the cancer registry (NCDR) and ONS mortality on the relative risk of cancer.

#### CPRD data only

The matched cohort described in **Section 6.3.4.1** was restricted to patients eligible to participate in the linkage scheme. No updates of cancer events were applied; all cancer events were identified from the CPRD.



**Incorporation of linked data: diagnosis identified in the CPRD OR cancer registry (concordant and discordant)**

Similarly to the CPRD only data analysis, the matched cohort described in **Section**

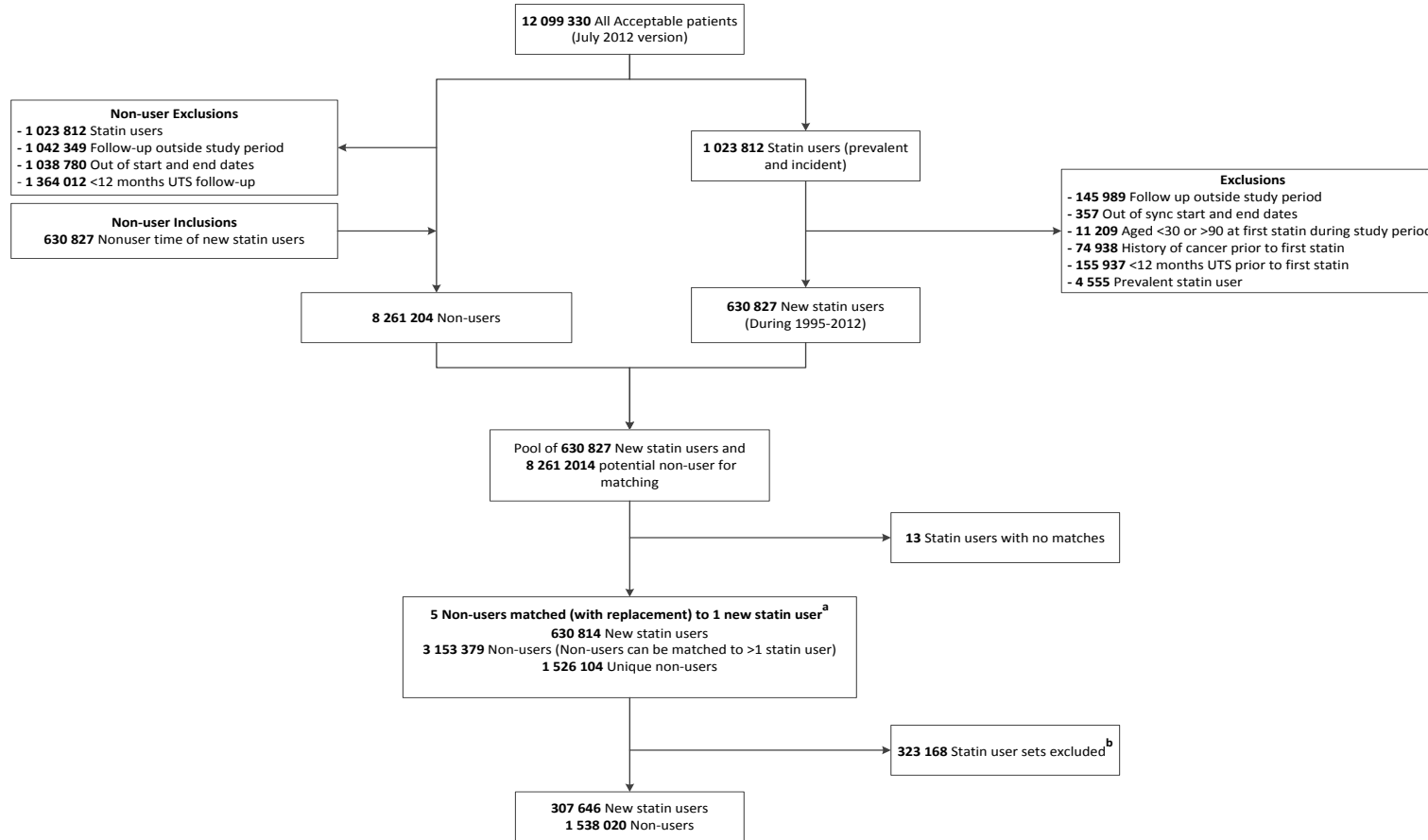
**6.3.4.1** was restricted to patients eligible to participate in the linkage scheme.

However, cancer outcomes identified in both the CPRD *OR* NCDR were included as events. For a concordant diagnosis recorded on different dates, the earlier of the two dates was assigned as the date of diagnosis. In addition, if the assigned NCDR date of diagnosis occurred before the statin treatment start date the statin user (and matched non-users) was excluded from further analyses.

**Incorporation of linked data: diagnosis identified in the CPRD AND cancer registry (concordant)**

For the concordant analysis, only concordant diagnoses were included as cancer events. Diagnoses identified in the CPRD alone and not in the NCDR were censored at the date of CPRD diagnosis and vice-versa.

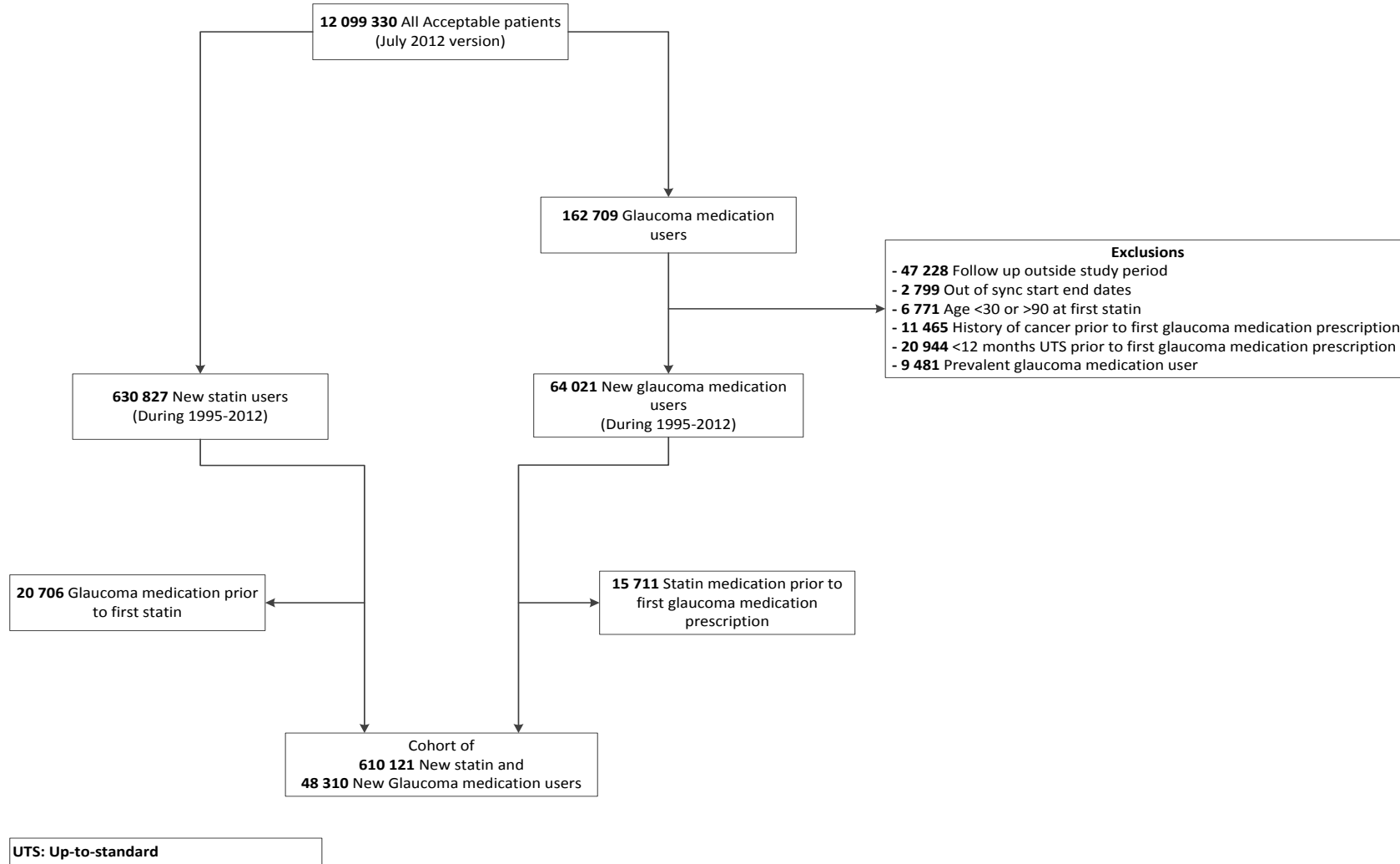
Figure 6.6: Flow diagram of inclusion exclusion of new statin users matched to non-users



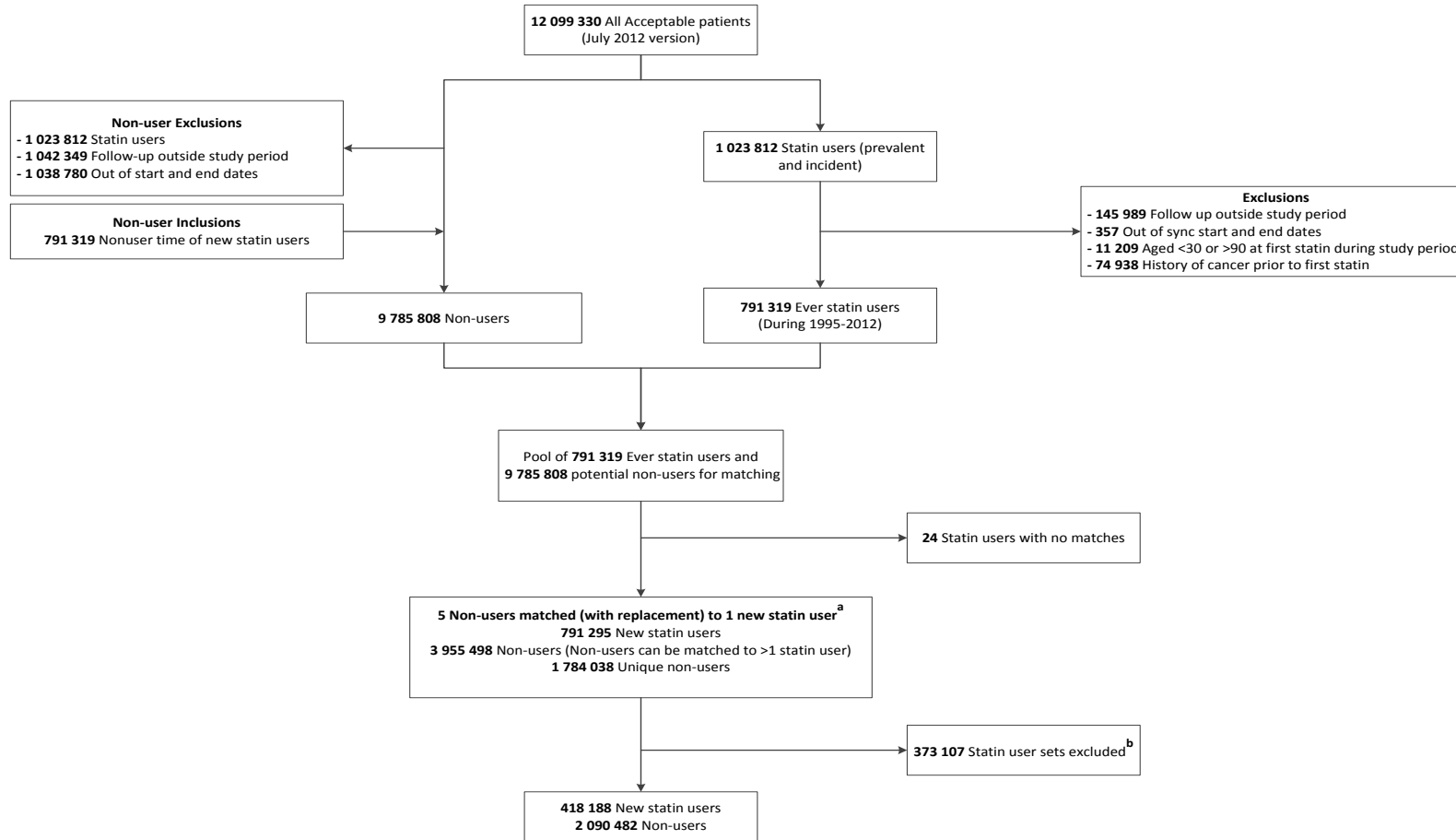
<sup>a</sup> Matched on treatment start date, age ( $\pm 2.5$ ), sex, general practice; 1:5 ratio of statin users to non-users.

<sup>b</sup> Intention to treat analysis: If a non-user went on to become a statin user, the matched set in which he/she was classified as a statin user was excluded, but the set in which he/she was classified as a non-user was included. UTS: Up-to-standard

**Figure 6.7: Flow diagram of inclusion exclusion of new statin and new glaucoma medication users**



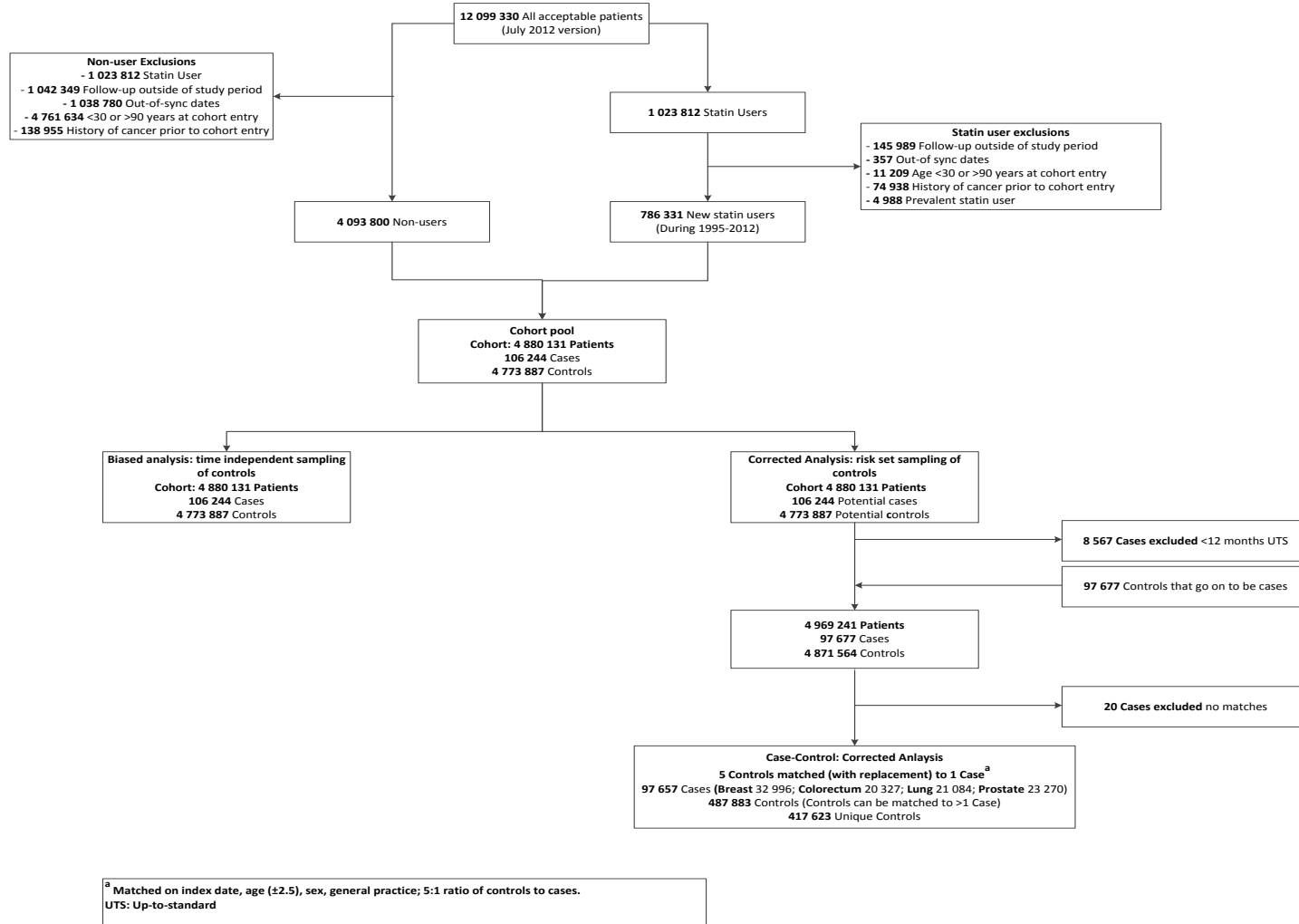
**Figure 6.8: Flow diagram of inclusion exclusion of statin users (incident and prevalent) matched to non-users**



<sup>a</sup> Matched on treatment start date, age ( $\pm 2.5$ ), sex, general practice; 1:5 Ratio of statin users to non-users.

<sup>b</sup> Intention to treat analysis: If a non-user went on to become a statin user, the matched set in which he/she was classified as a statin user was excluded, but the set in which he/she was classified as a non-user was included.  
UTS: Up-to-standard

**Figure 6.9: Flow diagram of the cohort of new statin users and non-users and a case-control design nested within the cohort**



## 6.4 Results: Descriptive Analysis

Overall, three cohort designs and one case-control design was used in this thesis, all of which comprised of statin users as the main treatment group.

### 6.4.1 *New statin and non-user cohort*

The matched cohort of *new statin* users and non-users consisted of 630,814 *new statin* users and 3,153,379 non-users. Overall, 630,484 (99.95%) statin users were matched to 5 non-users. All further analyses on this cohort were conducted on an intention-to-treat basis, which excluded 323,168 of the 630,814 statin users (as well as their non-user matches) leaving a total of 307,646 *new statin* users matched to 1,538,020 non-users (**Figure 6.6**). Overall, the majority of non-users were selected once as a match (837,025 non-users, 54.4%), while 478,776 non-users were matched on two occasions (31.1%), 169,671 non-users matched to three statin users (11.0%), and 52,548 (3.4%) non-users matched to 4-8 statin users.

**Table 6.1 (a)** presents the overall distribution of demographics between the two groups. Distributions of age and sex (matching factors) were identical between the two groups ( $p=1.000$ ). In terms of lifestyle factors, *new statin* users were more likely to be current or ex-smokers and have a higher BMI ( $>25$ ) compared to their counterpart non-users ( $p<0.001$ ). In addition, statin users were more likely to have a diagnosis of diabetes, coronary heart disease, hypertension, and hyperlipidaemia compared to non-users ( $p<0.001$ ). In terms of medications, statin users were also more likely to have been taking NSAIDs and antihypertensive medications ( $p<0.001$ ). Consultation visits to the GP were slightly higher (within 1 year of treatment start date) among statin users compared to non-users: mean consultation rate per year was 11 for statin users vs 6 for non-users ( $p<0.001$ ).

**Table 6.1: Demographics for the matched *new statin* cohort and nested case-control design (risk-set sampling)**

	(a) <i>New statin</i> vs matched non-users cohort				(b) Nested case-control				
	<i>New statin</i>		Non- <i>Statin</i>		Control		Case <sup>b</sup>		P-value <sup>a</sup>
<b>All Patients</b>	307646		1538020		487883		97657		0.999
<b>Age</b>									
30-39	14368	(4.7)	71841	(4.7)	5680	(1.2)	1136	(1.2)	
40-49	53773	(17.5)	268869	(17.5)	33565	(6.9)	6713	(6.9)	
50-59	91622	(29.8)	458113	(29.8)	78097	(16.0)	15619	(16.0)	
60-69	80903	(26.3)	404517	(26.3)	127639	(26.2)	25530	(26.1)	
70-79	46576	(15.1)	232884	(15.1)	142428	(29.2)	28479	(29.2)	
80+	20404	(6.6)	101796	(6.6)	100474	(20.6)	20180	(20.7)	
<b>Sex</b>									0.873
Male	163667	(53.2)	818167	(53.2)	233391	(47.8)	46744	(47.9)	
Female	143979	(46.8)	719853	(46.8)	254492	(52.2)	50913	(52.1)	
<b>Smoking status</b>									<0.001
Non	115053	(37.4)	689504	(44.8)	200898	(41.2)	37450	(38.3)	
Current	75132	(24.4)	333811	(21.7)	86533	(17.7)	20270	(20.8)	
Ex	115704	(37.6)	439600	(28.6)	166986	(34.2)	36894	(37.8)	
Unknown	1757	(0.6)	75105	(4.9)	33466	(6.9)	3043	(3.1)	
<b>BMI</b>									<0.001
<20	6494	(2.1)	63518	(4.1)	22965	(4.7)	5986	(6.1)	
20-25	65977	(21.4)	447459	(29.1)	146544	(30.0)	31148	(31.9)	
>25	218815	(71.1)	804975	(52.3)	259904	(53.3)	48307	(49.5)	
Unknown	16360	(5.3)	222068	(14.4)	58470	(12.0)	12216	(12.5)	

[Table 6.1 continued over]

[Table 6.1 continued]										
	(a) <i>New statin</i> vs matched non-users cohort					(b) Nested case-control				
	<i>New statin</i>		Non- <i>Statin</i>		P-value <sup>a</sup>	Control		Case <sup>b</sup>		P-value <sup>a</sup>
<b>Alcohol status</b>					<0.001					<0.001
<i>Non</i>	38768	(12.6)	167407	(10.9)		55058	(11.3)	9944	(10.2)	
<i>Ex</i>	12350	(4.0)	36426	(2.4)		21140	(4.3)	4674	(4.8)	
<i>Current</i>	8096	(2.6)	39943	(2.6)		11352	(2.3)	2276	(2.3)	
<i>rare&lt;2u/d</i>	56488	(18.4)	245204	(15.9)		92587	(19.0)	17844	(18.3)	
<i>moderate3-6u/d</i>	146339	(47.6)	709770	(46.1)		220934	(45.3)	44250	(45.3)	
<i>excessive &gt;6u/d</i>	29005	(9.4)	127556	(8.3)		32302	(6.6)	7601	(7.8)	
<i>Unknown</i>	16600	(5.4)	211714	(13.8)		54510	(11.2)	11068	(11.3)	
<b>Diabetes</b>	88714	(28.8)	105621	(6.9)	<0.001	74658	(15.3)	15420	(15.8)	<0.001
<b>CHD</b>	69974	(22.7)	58168	(3.8)	<0.001	58055	(11.9)	12011	(12.3)	<0.001
<b>Heart Failure</b>	11877	(3.9)	22732	(1.5)	<0.001	21111	(4.3)	5167	(5.3)	<0.001
<b>Hypertension</b>	148318	(48.2)	325190	(21.1)	<0.001	173647	(35.6)	34454	(35.3)	0.064
<b>Hyperlipidaemia</b>	99255	(32.3)	65564	(4.3)	<0.001	56159	(11.5)	10985	(11.2)	0.019
<b>NSAIDs/Aspirin</b>	124978	(40.6)	342029	(22.2)	<0.001	168153	(34.5)	37663	(38.6)	<0.001
<b>Antihypertensives</b>	168613	(54.8)	353061	(23.0)	<0.001	208667	(42.8)	44210	(45.3)	<0.001
<b>OC</b>	2484	(0.8)	16837	(1.1)	<0.001	4500	(0.9)	1123	(1.1)	<0.001
<b>HRT</b>	16550	(5.4)	78697	(5.1)	<0.001	20794	(4.3)	5078	(5.2)	<0.001
<b>Consultations</b>										
<b>Mean (SD)</b>	10.6	(8.9)	5.8	(7.3)	<0.001	7	(8.5)	11.8	(11.9)	<0.001

<sup>a</sup> P-values (two-sided) were from *t* tests (continuous factor) or chi-square test (categorical factor).  
<sup>b</sup> Includes all cases of breast, colorectal, lung, and prostate cancer



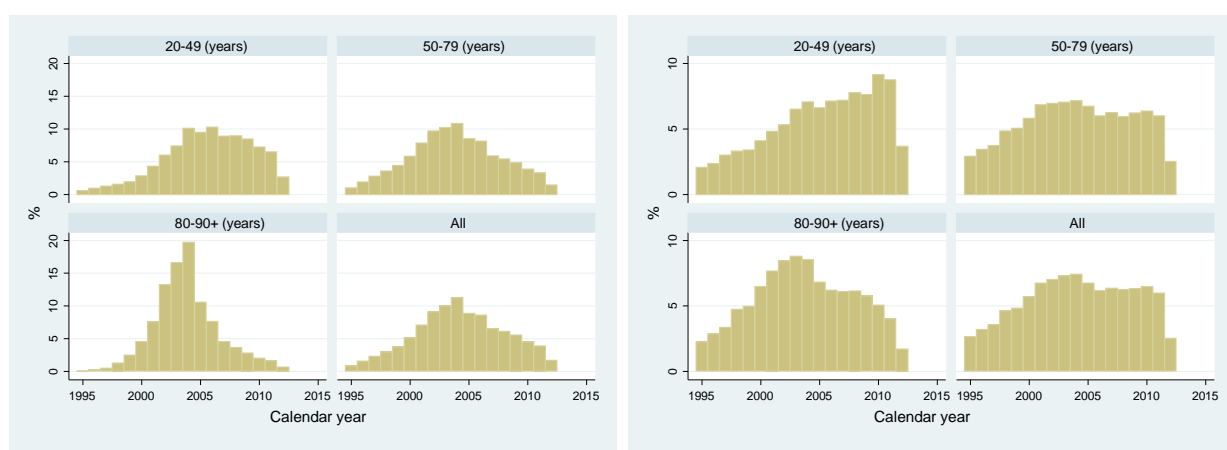
### 6.4.2 New statin users vs new glaucoma medication users

Overall, 630,827 new statin users were identified, of which 20,706 had a recorded prescription for glaucoma medications prior to the first statin prescription (Figure 6.7). In total, 162,709 patients with a recorded prescription for glaucoma related medications were identified; 98,688 patients were excluded as they did not meet the inclusion criteria leaving 64,021 new glaucoma medication users. Of the remaining 64,021 new glaucoma medication users, 15,711 had a previous statin prescription. Leaving a cohort of 48,310 glaucoma medication users and 610,121 new statin users (Figure 6.7).

Figure 6.10: Distribution of treatment start date by age and treatment group

(a) Statin Initiation

(b) Glaucoma Initiation



Appendix D, Table 11.1 (c) shows the distribution of demographics between new statin users and new glaucoma medication users. Overall age at treatment start date differed between the two groups ( $p < 0.001$ ). Higher proportions of statin initiators compared to glaucoma medication initiators were seen among ages 50-79. In contrast, a slightly higher proportion of glaucoma medication users were  $\geq 80$  years (statin 8.8% vs glaucoma 20.4%). Glaucoma medication users were less likely to smoke; lower proportions of non-smokers compared to statin users, in addition

to a lower proportion of current and ex-smokers ( $p<0.001$ ). BMI measures between 20-25 were higher among new glaucoma medication users compared to *new statin* users (32% vs 23%), However, a higher proportion of statin users had a BMI>25 compared to glaucoma medication users (70% vs 51%, overall  $p<0.001$ ). Glaucoma medication users were also less likely, compared to statin users, to have been diagnosed with diabetes, CHD, hypertension, or hyperlipidaemia. Use of NSAIDs or antihypertensive medications was also lower among glaucoma medication users. Consultation rates in the year prior to drug initiation was similar between the two groups (statin users, rate=10.7; glaucoma medication users, rate=9.2).

**Figure 6.10** depicts initiation of the two treatment groups (statin and glaucoma medication use) by calendar year. **Figure 6.10 (a)** shows the uptake of statin use from 1995-2012. A sharp increase in statin initiation can be seen between 1995 and 2004, a trend which later declines after 2004. Similarly, glaucoma medication users also show an increase in uptake from 1995 to 2004. Although not as sharp as with statin users, a small decline can be observed post-2004; however, uptake of glaucoma medications remained relatively constant from 2005 onwards.

#### **6.4.3 Ever statin user matched cohort**

Of the initial 1,023,812 patients identified as having a recorded prescription for a statin during the study period (1995-2012), 231,493 patients were excluded for failing to meet exclusion criteria (**Figure 6.8**). In total, 3,955,498 non-users were matched to 791,295 statin users (prevalent and incident). The final cohort included 2,090,482 non-users matched to 418,188 statin users – 111,172 extra prevalent statin users were added to the *new statin* cohort (**Figure 6.8**).

**Appendix D, Table 11.1 (a)** shows the demographics at treatment start date of the ever statin user cohort. Overall, similar distributions of most potential confounding factors were observed between ever statin users and *new statin* users (**Appendix D, Table 11.1 (b)**) Antihypertensive use was slightly higher among *new statin* users compared to prevalent users (*new statin*, 55%; prevalent statin, 44%). Similarly, the proportion for NSAID use was slightly higher among *new statin* users compared to prevalent statin users (*new statin*, 41%; prevalent statin, 32%).

#### **6.4.4 Nested case-control design**

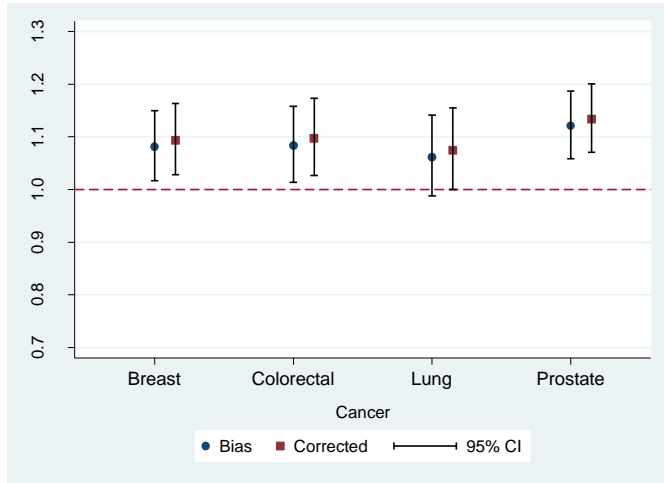
A cohort of *new statin* users (N=786,331) and non-users (N=4,093,800) aged 30-90 years at cohort entry was identified. From this cohort a total of 106,244 incident cancer (breast, colorectal, lung, and prostate) cases and 4,773,887 non-cases (controls) were identified (**Figure 6.9**). A matched case control design nested within this cohort was conducted. From the initial 106,244 cases identified, 8,567 cases were excluded because they had <12 months UTS follow-up prior to the index date. Control-time of the remaining 97,677 cases were included in the cohort for further consideration as potential controls. The final cohort consisted of 97,657 cases matched to 417,623 controls.

**Appendix D, Table 11.2** shows the demographics of the case-control (time-independent sampling) of *new statin* users and non-users. Cases in general were older compared to controls ( $p<0.001$ ). Lifestyle factors including smoking status, BMI, and alcohol status were similarly distributed between cases and controls ( $p<0.001$ ). Demographics from the matched nested case-control (risk-set sampling) showed an even distribution of age and sex compared to the time-independent

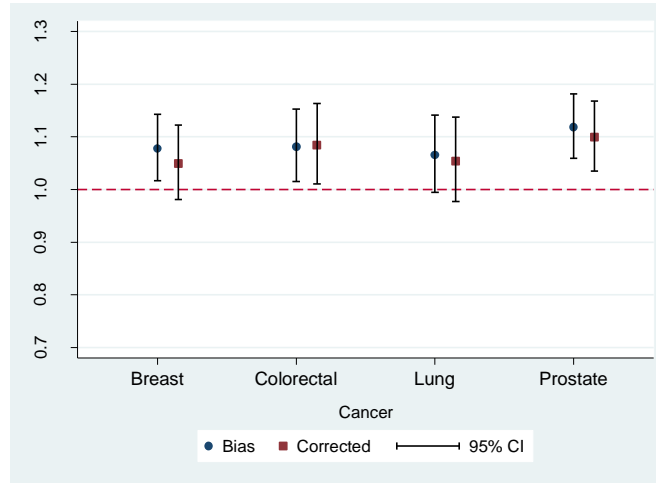
case-control design (**Table 6.1 (b)**). However, differences were observed between cases and controls in terms of lifestyle factors, co-morbidities and co-medications.

**Figure 6.11: Relative risk estimates and corresponding 95% confidence intervals for each bias analysis by cancer type**

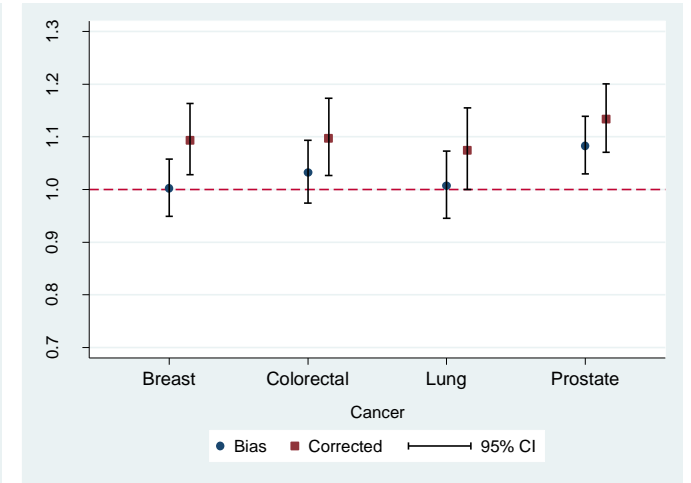
**(a) Immortal time bias (treatment definition: 2 statins)**



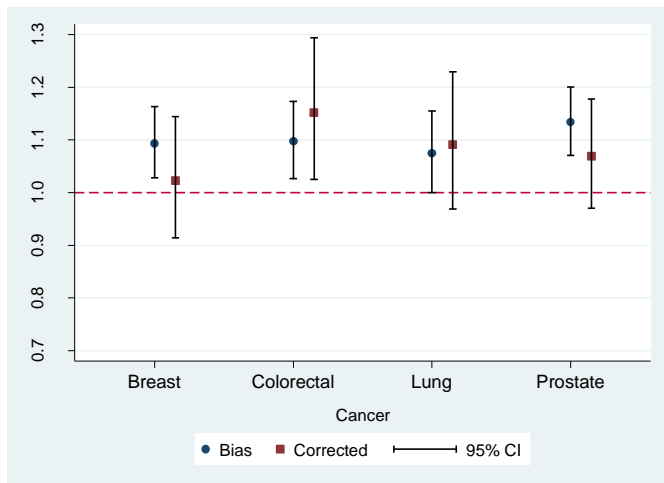
**(b) Protopathic bias**



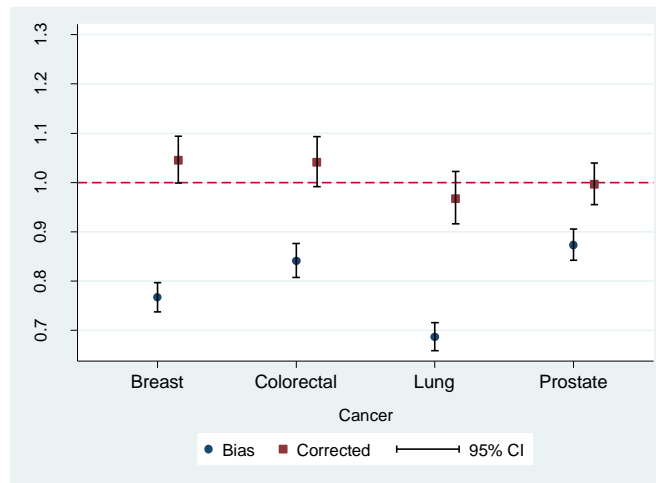
**(c) Prevalent user bias**



**(d) Healthy user bias**



**(e) Time-window bias**



## 6.5 Results: Impact of bias

### 6.5.1 Overview

Overall, the impact of four of the five biases in the context of the statin-cancer association was minimal. Only time-window bias showed a consistent and substantial impact, biasing the relative risk of cancer among statin users toward a protective effect.

### 6.5.2 Immortal time bias

**Table 6.2 and Figure 6.11 (a)** present relative risk estimates,  $\Delta\beta$  estimates, and corresponding 95% confidence intervals for both the biased and correct analyses of immortal time bias. Analyses that required two consecutive recorded statin prescriptions (**Table 6.2 (a)**) yielded  $\Delta\beta = -0.01$  (95% CI; ranging from -0.11, 0.09) across all four cancer types. Compared to the biased analysis, the corrected analysis excluded on average 0.12 PY (~44 days) of immortal time. Further analysis that extended the minimum period of follow-up to six months (**Table 6.2 (b)**) yielded lower relative risk estimates tending toward the null compared to the less restrictive 2-statin definition. More variability between the biased and corrected analysis was observed,  $\Delta\beta$  for the 6-month follow-up analysis ranging from -0.07 (95% CI; -0.15, 0.02) for breast cancer to -0.05 (95% CI; -0.13, 0.03) for prostate cancer.

For the 2-statin treatment definition, only corrected confidence interval estimates for lung cancer included 1. Breast cancer corrected analysis showed a marginal increased risk (RR=1.09; 95% CI; 1.03, 1.16). Prostate and colorectal cancer showed borderline increased risk in both biased and corrected analyses. For the 6-month treatment definition, confidence intervals from the corrected analyses included 1

for prostate cancer. For colorectal, lung, and prostate cancer, increased risks associated with the corrected design were observed (colorectal: RR=1.10; 95% CI; 1.03, 1.18; lung: RR=1.08; 95% CI; 1.01, 1.17; prostate: RR=1.11; 95% CI; 1.05, 1.18). Furthermore, risk estimates were marginally different between the biased (lower relative risk estimate) and corrected analysis (higher relative risk estimate). Biased and corrected risk estimates were generally consistent in terms of direction.

### **6.5.3 Protopathic bias**

The impact of protopathic bias on the statin-cancer association was minimal. **Table 6.3 and Figure 6.11 (b)** show relative risk estimates,  $\Delta\beta$  estimates, and corresponding 95% confidence intervals for both the biased and corrected analysis. Biased analyses incorporated a 0-lag period, in contrast to the corrected analysis which included a 360-day lag period. Protopathic bias  $\Delta\beta$  ranged from 0.00 (95% CI; -0.10, 0.09) for colorectal cancer to 0.03 (95% CI; -0.06, 0.12) for breast cancer.

Relative risk point estimates showed little or no change from biased to corrected analysis. Of note, confidence intervals for corrected protopathic bias analyses spanned 1 for breast and lung cancer.

**Table 6.2: Immortal time bias relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Relative Risk <sup>a</sup> (95% CI)	$\Delta\beta^b$ (95% CI)
<b>(a) Minimum of 2 statin prescriptions</b>						
<b>Breast Cancer</b>						
Biased	Unexposed	505 031	5.94	8254 (1.6)	1.08	-0.01
	Exposed	117 691	5.88	2154 (1.8)	(1.02, 1.15)	(-0.10, 0.08)
Corrected	Unexposed	502 829	5.85	8149 (1.6)	1.09	
	Exposed	117 691	5.77	2154 (1.8)	(1.03, 1.16)	
<b>Colorectal Cancer</b>						
Biased	Unexposed	1 035 532	5.94	7061 (0.7)	1.08	-0.01
	Exposed	251 556	5.83	1787 (0.7)	(1.01, 1.16)	(-0.11, 0.08)
Corrected	Unexposed	1 030 623	5.85	6980 (0.7)	1.10	
	Exposed	251 556	5.71	1787 (0.7)	(1.03, 1.17)	
<b>Lung Cancer</b>						
Biased	Unexposed	1 035 532	5.94	7145 (0.7)	1.06	-0.01
	Exposed	251 556	5.83	1931 (0.8)	(0.99, 1.14)	(-0.11, 0.09)
Corrected	Unexposed	1 030 623	5.85	7082 (0.7)	1.07	
	Exposed	251 556	5.71	1931 (0.8)	(1.00, 1.16)	
<b>Prostate Cancer</b>						
Biased	Unexposed	530 501	5.94	9686 (1.8)	1.12	-0.01
	Exposed	133 865	5.78	2517 (1.9)	(1.06, 1.19)	(-0.09, 0.07)
Corrected	Unexposed	527 794	5.85	9606 (1.8)	1.13	
	Exposed	133 865	5.66	2517 (1.9)	(1.07, 1.20)	
<b>(b) Minimum of 6 months follow-up</b>						
<b>Breast Cancer</b>						
Biased	Unexposed	488 154	6.04	8134 (1.7)	1.01	-0.07
	Exposed	113 735	6.06	2019 (1.8)	(0.95, 1.07)	(-0.15, 0.02)
Corrected	Unexposed	478 769	5.65	7644 (1.6)	1.08	
	Exposed	113 735	5.56	2019 (1.8)	(1.01, 1.15)	
<b>Colorectal Cancer</b>						
Biased	Unexposed	1 000 777	6.05	6929 (0.7)	1.04	-0.06
	Exposed	242 986	6.02	1725 (0.7)	(0.97, 1.11)	(-0.15, 0.04)
Corrected	Unexposed	980 554	5.65	6606 (0.7)	1.10	
	Exposed	242 986	5.52	1725 (0.7)	(1.03, 1.18)	
<b>Lung Cancer</b>						
Biased	Unexposed	1 000 777	6.05	7040 (0.7)	1.02	-0.06
	Exposed	242 986	6.02	1869 (0.8)	(0.95, 1.10)	(-0.16, 0.04)
Corrected	Unexposed	980 554	5.65	6735 (0.7)	1.08	
	Exposed	242 986	5.52	1869 (0.8)	(1.01, 1.17)	
<b>Prostate Cancer</b>						
Biased	Unexposed	512 623	6.05	9496 (1.9)	1.06	-0.05
	Exposed	129 251	5.98	2403 (1.9)	(1.00, 1.12)	(-0.13, 0.03)
Corrected	Unexposed	501 785	5.66	9102 (1.8)	1.11	
	Exposed	129 251	5.48	2403 (1.9)	(1.05, 1.18)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates



**Table 6.3: Protopathic bias relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Relative Risk <sup>a</sup> (95% CI)	$\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Biased (0-day lag)	Unexposed	553 656	5.81	9 000 (1.6)	1.08	0.03
	Exposed	131 581	5.74	2 377 (1.8)	(1.02, 1.14)	(-0.06, 0.12)
Corrected (360-day lag)	Unexposed	434 616	5.52	6 932 (1.6)	1.05	
	Exposed	107 399	5.50	1 888 (1.8)	(0.98, 1.12)	
<b>Colorectal Cancer</b>						
Biased (0-day lag)	Unexposed	1 131 970	5.79	7 641 (0.7)	1.08	0.00
	Exposed	281 347	5.67	1 948 (0.7)	(1.01, 1.15)	(-0.10, 0.09)
Corrected (360-day lag)	Unexposed	895 020	5.51	6 136 (0.7)	1.08	
	Exposed	231 466	5.45	1 679 (0.7)	(1.01, 1.16)	
<b>Lung Cancer</b>						
Biased (0-day lag)	Unexposed	1 131 970	5.79	7 743 (0.7)	1.07	0.01
	Exposed	281 347	5.67	2 119 (0.8)	(0.99, 1.14)	(-0.09, 0.11)
Corrected (360-day lag)	Unexposed	895 020	5.51	6 325 (0.7)	1.05	
	Exposed	231 466	5.45	1 797 (0.8)	(0.98, 1.14)	
<b>Prostate Cancer</b>						
Biased (0-day lag)	Unexposed	578 314	5.77	10 417 (1.8)	1.12	0.02
	Exposed	149 766	5.60	2 726 (1.8)	(1.06, 1.18)	(-0.06, 0.10)
Corrected (360-day lag)	Unexposed	460 404	5.51	8 537 (1.9)	1.10	
	Exposed	124 067	5.41	2 336 (1.9)	(1.04, 1.17)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup> $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

#### 6.5.4 Prevalent user bias

**Table 6.4 and Figure 6.11 (c)** present relative risk estimates,  $\Delta\beta$  estimates, and corresponding 95% confidence intervals for prevalent user bias. Prevalent user  $\Delta\beta$  ranged from -0.09 (95% CI; -0.17, -0.01) for breast cancer to -0.05 (95% CI; -0.12, 0.03) for prostate cancer.

For all four cancer types, the biased analysis, which included prevalent statin users, yielded relative risk estimates that were consistently lower compared to the new user analysis (**Table 6.4 and Figure 6.11 (c)**). Confidence intervals for the corrected analysis included 1 only for lung cancer. Of note, prevalent users of statins represented more than a third of total statin use during the study period (1995-2012): women only (36%), men and women (38%), and men (46%).

#### 6.5.5 Healthy user bias

**Table 6.5 and Figure 6.11 (d)** present relative risk estimates,  $\Delta\beta$  estimates, and corresponding 95% confidence intervals for the biased and corrected analysis of healthy user bias. The impact of healthy user bias varied by cancer type:  $\Delta\beta$  ranged from -0.05 (95% CI; -0.18, 0.09) for colorectal cancer to 0.07 (95% CI; -0.06, 0.19) for breast cancer.

No consistent pattern was observed in terms of direction of relative risk estimates between biased and corrected analysis among the four cancer types. Breast (RR=1.02; 95% CI; 0.91, 1.14) and prostate (RR=1.07; 95% CI; 0.97, 1.18) cancer yielded the lowest corrected relative risk estimates, while higher corrected relative risk estimates were observed for colorectal (RR=1.15; 95% CI; 1.02, 1.29) and lung cancer (RR=1.07; 95% CI; 0.97, 1.18). All corrected analyses confidence intervals

included 1 except for colorectal cancer. However, confidence intervals in the corrected analysis which compared statin users and glaucoma medication users were slightly larger due to the relatively smaller sample size.

### **6.5.6 Time-window bias**

**Table 6.6 and Figure 6.11 (e)** show relative risk estimates,  $\Delta\beta$  estimates, and corresponding 95% confidence intervals for both the biased and corrected analysis of time-window bias. Time-window bias yielded the most variability in terms of  $\Delta\beta$ , which ranged from -0.34 (95% CI; -0.41, -0.27) for lung cancer to -0.13 (95% CI; -0.19, -0.08) for prostate cancer.

Biased analyses yielded statistically significant protective effects across all four cancer types: relative risk estimates ranging from RR=0.69 (95% CI, 0.66, 0.72) for lung cancer to RR=0.87 (95% CI, 0.84, 0.91) for prostate cancer. Corrected analyses showed no association for all four cancer types: relative risk estimates ranged from RR=0.97 (95% CI, 0.92, 1.02) for lung cancer to RR=1.05 (95% CI, 1.00, 1.10) for breast cancer.

**Table 6.4: Prevalent user bias relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Relative Risk <sup>a</sup> (95% CI)	$\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Biased (Prevalent user)	Unexposed	812 670	5.31	12 520 (1.5)	1.00	-0.09
	Exposed	169 619	5.28	2 837 (1.7)	(0.95, 1.06)	(-0.17, -0.01)
Corrected (New user)	Unexposed	502 829	5.85	8 149 (1.6)	1.09	
	Exposed	117 691	5.77	2 154 (1.8)	(1.03, 1.16)	
<b>Colorectal Cancer</b>						
Biased (Prevalent user)	Unexposed	1 690 276	5.28	10 813 (0.6)	1.03	-0.06
	Exposed	369 963	5.17	2 475 (0.7)	(0.97, 1.09)	(-0.15, 0.03)
Corrected (New user)	Unexposed	1 030 623	5.85	6 980 (0.7)	1.10	
	Exposed	251 556	5.71	1 787 (0.7)	(1.03, 1.17)	
<b>Lung Cancer</b>						
Biased (Prevalent user)	Unexposed	1 690 276	5.28	10 709 (0.6)	1.01	-0.07
	Exposed	369 963	5.17	2 636 (0.7)	(0.95, 1.07)	(-0.16, 0.03)
Corrected (New user)	Unexposed	1 030 623	5.85	7 082 (0.7)	1.07	
	Exposed	251 556	5.71	1 931 (0.8)	(1.00, 1.16)	
<b>Prostate Cancer</b>						
Biased (Prevalent user)	Unexposed	877 606	5.26	14 579 (1.7)	1.08	-0.05
	Exposed	200 344	5.07	3 380 (1.7)	(1.03, 1.14)	(-0.12, 0.03)
Corrected (New user)	Unexposed	527 794	5.85	9 606 (1.8)	1.13	
	Exposed	133 865	5.66	2 517 (1.9)	(1.07, 1.20)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 6.5: Healthy user bias relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Relative Risk <sup>a</sup> (95% CI)	$\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Biased (non-user comparison group)	Unexposed	502 829	5.85	8 149 (1.6)	1.09	0.07
	Exposed	117 691	5.77	2 154 (1.8)	(1.03, 1.16)	(-0.06, 0.19)
Corrected (glaucoma medication comparison group)	Unexposed	21 634	4.96	381 (1.8)	1.02	
	Exposed	254 826	4.95	4 210 (1.7)	(0.91, 1.14)	
<b>Colorectal Cancer</b>						
Biased (non-user comparison group)	Unexposed	1 030 623	5.85	6 980 (0.7)	1.10	-0.05
	Exposed	251 556	5.71	1 787 (0.7)	(1.03, 1.17)	(-0.18, 0.09)
Corrected (glaucoma medication comparison group)	Unexposed	40 538	4.90	354 (0.9)	1.15	
	Exposed	561 295	4.83	4 249 (0.8)	(1.02, 1.29)	
<b>Lung Cancer</b>						
Biased (non-user comparison group)	Unexposed	1 030 623	5.85	7 082 (0.7)	1.07	-0.02
	Exposed	251 556	5.71	1 931 (0.8)	(1.00, 1.16)	(-0.15, 0.12)
Corrected (glaucoma medication comparison group)	Unexposed	40 538	4.90	331 (0.8)	1.09	
	Exposed	561 295	4.83	4 414 (0.8)	(0.97, 1.23)	
<b>Prostate Cancer</b>						
Biased (non-user comparison group)	Unexposed	530 501	5.94	9 686 (1.8)	1.12	0.06
	Exposed	133 865	5.78	2 517 (1.9)	(1.06, 1.19)	(-0.05, 0.17)
Corrected (glaucoma medication comparison group)	Unexposed	18 904	4.82	525 (2.8)	1.07	
	Exposed	306 469	4.73	5 984 (2.0)	(0.97, 1.18)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup> $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 6.6: Time-window bias relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Case status	N	Median Follow-up (years)	Statin user (%)	Relative Risk (95% CI)	$\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Biased (Time independent sampling)	Control	1 859 617	6.65	313 766 (16.9)	0.77	-0.31
	Case	30 283	5.46	5 096 (16.8)	(0.74, .80)	(-0.37, -0.25)
Corrected (Risk-set sampling)	Control	122 015	6.10	20 400 (16.7)	1.05	
	Case	27 965	6.00	4 955 (17.7)	(1.00, 1.09)	
<b>Colorectal Cancer</b>						
Biased (Time independent sampling)	Control	3 602 729	6.43	703 478 (19.5)	0.84	-0.21
	Case	17 753	6.20	5 195 (29.3)	(0.81, 0.88)	(-0.28, -0.15)
Corrected (Risk-set sampling)	Control	71 522	6.78	20 641 (28.9)	1.04	
	Case	16 689	6.64	5 111 (30.6)	(0.99, 1.09)	
<b>Lung Cancer</b>						
Biased (Time independent sampling)	Control	3 603 001	6.43	703 351 (19.5)	0.69	-0.34
	Case	17 481	6.01	5 322 (30.4)	(0.66, 0.72)	(-0.41, -0.27)
Corrected (Risk-set sampling)	Control	71 458	6.64	21 139 (29.6)	0.97	
	Case	16 459	6.40	5 221 (31.7)	(0.92, 1.02)	
<b>Prostate Cancer</b>						
Biased (Time independent sampling)	Control	1 709 268	6.23	382 437 (22.4)	0.87	-0.13
	Case	21 314	6.59	7 374 (34.6)	(0.84, 0.91)	(-0.19, -0.08)
Corrected (Risk-set sampling)	Control	85 660	7.20	29 489 (34.4)	1.00	
	Case	20 064	7.03	7 224 (36.0)	(0.96, 1.04)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup> $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

## **6.5.7 Sensitivity analyses: weighting, missing data, and censoring of treatment change**

### **6.5.7.1 Matching with replacement: down-weighting non-users**

Marginally lower relative risk estimates across all cancer types and bias analysis (immortal, protopathic, and prevalent user bias) were observed when applying a weighted analysis, down-weighting matched non-users that were matched on >1 occasion (**Appendix D, Tables 11.3-11.5**). Confidence intervals were of similar width suggesting relatively minimal influence of matching with replacement on effect estimates. In addition,  $\Delta\beta$  estimates from weighted analyses were similar to primary analyses estimates across all cancer types.

### **6.5.7.2 Missing data**

Differential proportions of missing data (unknown value) were observed for BMI, alcohol status, and smoking status between *new statin* users and matched non-users (**Table 6.1 (a)**). Sensitivity analyses assessed the impact of missing data and yielded similar relative risk estimates for imputed and missing category analyses compared to the primary complete case analysis (**Appendix D, Tables 11.6-11.10**). In addition,  $\Delta\beta$  estimates from both imputed and missing category analyses were similar to primary analyses estimates across all cancer types.

### **6.5.7.3 Censoring follow-up at treatment change**

**Appendix D, Tables 11.11-11.13** present relative risk estimates from analyses censoring follow-up at treatment switch. For example, non-user follow-up time was censored when a first statin prescription was recorded during follow-up. Similarly, statin users follow-up time was censored when statin use was stopped for a continuous period of 6-months. Higher relative risk estimates were observed for all

analyses censoring follow-up at treatment switch. Possible reasons for the increase in relative risk might be explained by the higher proportion of exposed cases relative to non-user cases compared to primary analysis proportions. In addition, unmeasured confounding may have also increased relative risk estimates.

## **6.6 Results: Impact of alternative outcome definitions**

### **6.6.1 Case definitions**

In comparison to the standard case definition, which required one definite malignant diagnosis code, higher relative risk estimates from the broader case definition were observed across all cancer types. However, the impact of alternative case definitions was minimal:  $\Delta\beta$  ranged from -0.03 for lung (95% CI; -0.13, 0.07) and prostate (95% CI; -0.11, 0.05) cancer to -0.01 for breast (95% CI; -0.10, 0.08) and colorectal cancer (95% CI; -0.10, 0.08) (**Table 6.7**). The majority of additional cases were identified for lung (increase of 717 unexposed and 265 exposed cases) and prostate cancer (increase of 832 unexposed and 327 exposed cases). These two cancer types yielded the most variability between relative risk estimates: lung and prostate cancer  $\Delta\beta = -0.03$ . The proportion of events remained similar between exposed and unexposed groups, even when patients from probable and possible diagnostic groups were included, suggesting no differential early cancer detection between statin users and non-users.

### **6.6.2 Linkage to the cancer registry**

Overall, the incorporation of patient-level linked data from the NCDR generated similar results to those that used only CPRD data restricted to linkage eligible patients and practices (**Table 6.8**).  $\Delta\beta$  due to incorporating linked data (outcomes from all data sources, CPRD OR NCDR) ranged from 0.01 (95% CI; -0.09, 0.11) for



prostate cancer to 0.05 (95% CI; -0.06, 0.16) for colorectal cancer (**Table 6.8**).

Similarly, analyses restricted to concordant diagnosis between the CPRD *AND* NCDR resulted in minimal impact on the statin-cancer association compared to analyses incorporating the CPRD alone.  $\Delta\beta$  estimates when incorporating linked data (only concordant diagnoses, CPRD *AND* NCDR) ranged from -0.01 (95% CI; -0.11, 0.11) for breast cancer to 0.03 (95% CI; -0.11, 0.17) for lung cancer. Of all four cancer types, the most variability in terms of relative risk estimates for each linkage analysis was observed for colorectal cancer. In comparison to relative risk estimates from linked data analysis (*AND* or *OR*), slightly higher relative risk estimates were observed when the CPRD alone was used.

**Table 6.7: Case definitions relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (%)	No. of outcomes (%)	Relative Risk <sup>a</sup> (95% CI)	$\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Standard case definition	Unexposed	502 829	5.85	8 149 (1.6)	1.09	-0.01
	Exposed	117 691	5.77	2 154 (1.8)	(1.03, 1.16)	(-0.10, 0.08)
Broad case definition	Unexposed	502 829	5.83	8 342 (1.7)	1.11	
	Exposed	117 691	5.75	2 222 (1.9)	(1.04, 1.18)	
<b>Colorectal Cancer</b>						
Standard case definition	Unexposed	1 030 623	5.85	6 980 (0.7)	1.10	-0.01
	Exposed	251 556	5.71	1 787 (0.7)	(1.03, 1.17)	(-0.10, 0.08)
Broad case definition	Unexposed	1 030 623	5.84	7 149 (0.7)	1.11	
	Exposed	251 556	5.70	1 850 (0.7)	(1.04, 1.18)	
<b>Lung Cancer</b>						
Standard case definition	Unexposed	1 030 623	5.85	7 082 (0.7)	1.07	-0.03
	Exposed	251 556	5.71	1 931 (0.8)	(1.00, 1.16)	(-0.13, 0.07)
Broad case definition	Unexposed	1 030 623	5.84	7 799 (0.8)	1.10	
	Exposed	251 556	5.70	2 196 (0.9)	(1.03, 1.18)	
<b>Prostate Cancer</b>						
Standard case definition	Unexposed	527 794	5.85	9 606 (1.8)	1.13	-0.03
	Exposed	133 865	5.66	2 517 (1.9)	(1.07, 1.20)	(-0.11, 0.05)
Broad case definition	Unexposed	527 794	5.84	10 438 (2.0)	1.17	
	Exposed	133 865	5.64	2 844 (2.1)	(1.11, 1.24)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup> $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 6.8: Linked data relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Relative Risk <sup>b</sup> (95% CI)	$\Delta\beta^c$ (95% CI)
<b>Breast Cancer</b>						
CPRD data only <sup>a</sup>	Unexposed	306 529	5.85	5 055 (1.6)	1.08	
	Exposed	71 631	5.80	1 336 (1.9)	(1.00, 1.17)	
CPRD <b>OR</b> linked NCDR and death registry data	Unexposed	304 987	5.85	5 484 (1.8)	1.06	0.02
	Exposed	71 456	5.80	1 441 (2.0)	(0.98, 1.14)	(-0.09, 0.13)
CPRD <b>AND</b> linked NCDR and death registry data	Unexposed	304 987	5.85	3 877 (1.3)	1.09	-0.01
	Exposed	71 456	5.80	1 029 (1.4)	(1.00, 1.19)	(-0.13, 0.11)
<b>Colorectal Cancer</b>						
CPRD data only <sup>a</sup>	Unexposed	633 100	5.88	4 267 (0.7)	1.16	
	Exposed	154 039	5.77	1 119 (0.7)	(1.07, 1.26)	
CPRD <b>OR</b> linked NCDR and death registry data	Unexposed	630 217	5.88	5 595 (0.9)	1.11	0.05
	Exposed	153 704	5.77	1 435 (0.9)	(1.03, 1.19)	(-0.06, 0.16)
CPRD <b>AND</b> linked NCDR and death registry data	Unexposed	630 217	5.88	3 441 (0.5)	1.14	0.02
	Exposed	153 704	5.77	897 (0.6)	(1.04, 1.26)	(-0.11, 0.15)
<b>Lung Cancer</b>						
CPRD data only <sup>a</sup>	Unexposed	633 100	5.88	4 276 (0.7)	1.06	
	Exposed	154 039	5.77	1 145 (0.7)	(0.97, 1.16)	
CPRD <b>OR</b> linked NCDR and death registry data	Unexposed	630 217	5.88	5 904 (0.9)	1.04	0.02
	Exposed	153 704	5.77	1 586 (1.0)	(0.96, 1.13)	(-0.11, 0.14)
CPRD <b>AND</b> linked NCDR and death registry data	Unexposed	630 217	5.88	3 646 (0.6)	1.03	0.03
	Exposed	153 704	5.77	966 (0.6)	(0.93, 1.14)	(-0.11, 0.17)
<b>Prostate Cancer</b>						
CPRD data only <sup>a</sup>	Unexposed	326 571	5.90	6 226 (1.9)	1.09	
	Exposed	82 408	5.74	1 578 (1.9)	(1.01, 1.17)	
CPRD <b>OR</b> linked NCDR and death registry data	Unexposed	325 230	5.90	7 039 (2.2)	1.08	0.01
	Exposed	82 248	5.74	1 810 (2.2)	(1.01, 1.16)	(-0.09, 0.11)
CPRD <b>AND</b> linked NCDR and death registry data	Unexposed	325 230	5.90	4 671 (1.4)	1.09	0.00
	Exposed	82 248	5.74	1 192 (1.4)	(1.00, 1.18)	(-0.11, 0.11)

<sup>a</sup> Restricted to CPRD linkage eligible patients;

<sup>b</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>c</sup>  $\Delta\beta$ = Difference between "CPRD data only" and "linked" log relative risk estimates

## **6.7 Discussion**

### **6.7.1 Overview**

In this chapter, the impact of several potential drivers of discrepant results was systematically evaluated. Potential drivers included study design bias, different case definitions, and data linkage. These potential drivers have often been cited as potential reasons for conflicting findings between past observational studies, and also have the potential to influence future studies.<sup>69, 70, 116, 118, 119, 187</sup>

The statin-cancer association was selected as a basis to evaluate the impact of these potential drivers. Findings from past RCTs<sup>40, 41, 50, 78, 198</sup> and pharmacoepidemiological studies<sup>131, 206, 207</sup> lead to the assumption that no causal link exists between statin use and cancer risk, and thus a “corrected” analysis should yield confidence interval estimates that include 1. Results observed in this chapter showed that six of the seven potential drivers had minimal effect on the overall conclusions of an example study examining statin use and cancer risk. On the other hand, the findings demonstrated how a single bias (time-window bias) can influence quantitative findings of a study if not mitigated appropriately.

### **6.7.2 Impact of bias on the statin-cancer association**

In the context of the statin-cancer association, the impact of four of the five biases examined was minimal. Only time-window bias consistently drove the estimated statin-cancer association toward a protective effect. Immortal time, protopathic, prevalent user, and healthy user bias had minimal impact on the statin-cancer association.

### **6.7.2.1 Impact of time-window bias**

Time-window bias showed the greatest impact among all biases examined in this study. A consistent change from a protective effect (biased analysis) to a null association (corrected analysis) was observed across all four cancer types. Similar effects were observed to those presented by Suissa *et al.*<sup>119</sup> who also examined lung cancer risk among statin users.

### **6.7.2.2 Impact of immortal time bias**

The overall impact of immortal time bias in the context of this study was marginal. Point estimates of the direction of bias (biased vs corrected analyses) was consistently towards a lower relative risk across all four cancer types, although direction of  $\Delta\beta$  was uncertain: 95% CIs including negative and positive  $\Delta\beta$  estimates. Only confidence intervals corresponding to the corrected analysis for lung cancer spanned 1. Borderline confidence intervals were observed for the other three cancer types; the most pronounced effect observed was for prostate cancer.

These findings contrast with those from past pharmacoepidemiological studies that might have been affected by immortal time bias.<sup>69</sup> Importantly, two main components have been shown to drive the extent of this bias: (i) the proportion of immortal time relative to total exposed time; and (ii) the ratio of person time between exposed and unexposed groups.<sup>69</sup> The relatively minimal impact of immortal time in this study can be partly explained by the low proportion of immortal time relative to the exposed person time (immortal time/total exposed time). Estimates ranged from 2% ((706491-693997)/706491) for treatment definition 1 to 9% ((608213-554289)/608213) for treatment definition 2, which was not unexpected since the treatment definitions implemented were conservative

and representative of definitions utilised by most observational statin-cancer studies (**Chapter 3**). Furthermore, the ratio of observed person-years between unexposed and exposed was moderate (~4.2 across all cancer types) – a lower number of non-users relative to statin users would increase immortal time. These two measures were relatively low/high in comparison to figures reported by Suissa *et al.*<sup>69</sup> The ratio of immortal time to total exposed time was 103% (316.5/308.1) (high proportion of immortal time among exposed). The ratio of person-years between exposed and unexposed was 0.8 (lower number of unexposed patients compared to exposed) which could partly explain the contrasting minimal impact of immortal time bias observed in this study.

### **6.7.2.3 Impact of protopathic bias**

The impact of protopathic bias was minimal, with only breast and prostate cancer showing slight variability between biased and correct analyses. Although the overall impact of protopathic bias was relatively small, “corrected” relative risk estimates were less than or equal to “biased” estimates when no lag period was implemented. Based on the systematic review conducted in **Chapter 3** which examined statin-cancer pharmacoepidemiological studies, most studies implemented a lag-time across all cancer types to mitigate protopathic bias. Despite most studies implementing a lag period, the impact of protopathic bias in the statin-cancer setting may be minimal because there may not be a common pre-diagnosis cancer symptom(s) that would typically lead to a statin prescription. However, other situations may have a larger impact; for example, discontinuation of statin use in a case-control setting or symptoms such as cough treatment, gastro-intestinal problems, and pain.

#### **6.7.2.4 Impact of prevalent user bias**

Relative risk estimates across all cancer types differed marginally when considering *new statin* users compared to the inclusion of both incident and prevalent statin users. However, the inclusion of prevalent statin users yielded consistently lower, albeit marginal, relative risk estimates compared to results from the new user analysis, suggesting an effect tending toward the null among prevalent statin users. The impact between a new user and prevalent user analysis observed in this study was similar to that observed in a past study by Schneeweiss *et al.*<sup>208</sup> who investigated 1-year mortality among elderly patients in Pennsylvania, USA. Similar to this study, a small difference in risk was observed by Schneeweiss *et al.*<sup>208</sup>  $\Delta\beta = -0.03$ . Over half of all statin users in this study were incident users, in contrast to that of Schneeweiss *et al.*<sup>208</sup> where less than half of all statin users were new users. This may have brought about slight differences in results observed in this chapter in terms of power and sample size compared to Schneeweiss *et al.*<sup>208</sup> The proportion of new and prevalent users of any such drug is dependent on the prevalence of the drug and the start point (treatment start date) at which new users are defined.

#### **6.7.2.5 Impact of healthy user bias**

Overall, the impact of healthy user bias was minimal, and corrected analyses confidence intervals crossed 1 for all cancer types except colorectal cancer. Healthy user analyses (corrected vs biased) showed the most variability within cancer type. This may have been caused by the lack of power in the study, or by a spurious association between glaucoma medication users and the risk of cancer; although, previous literature suggests otherwise for the latter.<sup>209, 210</sup> Previous studies have argued that statin users are possibly healthier compared to their non-user

counterparts from the general population due to the health seeking behaviour of a patient receiving a preventative drug.<sup>118</sup> At inception of statin use, descriptive analysis examining lifestyle factors, co-morbidities and medications suggest the contrary. A greater proportion of statin users were overweight, current smokers, with more co-morbidities compared to non-users. However, visits to the GP were generally higher among statin users compared to their counterpart non-users (11 vs 6 visits per year). “Corrected” analysis relative risk estimates for both breast and prostate cancer moved toward the null, which could be partly explained by minimisation of detection bias. Although “corrected” relative risk estimates moved upward from the null for colorectal and lung cancer, which may be due to fatality of disease.

### **6.7.3 Sensitivity analysis**

Three analytical aspects of the implemented study designs were examined in further sensitivity analyses: replacement of non-users in the matching process; missing data; and censoring patients at treatment switch. Overall, study findings were robust in terms of replacement of non-users and missing data. However, increased relative risk estimates were observed when censoring at treatment switch, in comparison to the primary analysis which utilised an intention to treat design. In the censored analysis, slightly higher proportions of exposed cases relative to non-users were observed compared to the primary analysis which could have contributed to the increased cancer risk associated with statin use. However, unmeasured confounding could have also influenced effect estimates. That being said,  $\Delta\beta$  estimates were similar between primary ITT analyses compared to sensitivity analyses incorporating treatment switch.



#### **6.7.4 Impact of outcome definition on the statin-cancer association**

Similar to the impact of bias, the implementation of alternative approaches to define cancer outcomes did not influence study findings substantially.

##### **6.7.4.1 Case definition**

Using a broader case definition to identify cancer events (addition of patients from possible and probable diagnostic groups) had relatively minimal impact on the effect of statin use on cancer risk. This was in part due to the relatively low number of additional cases included in the broad definition. The relatively low number of additional cases included in the broad definition can be explained by the majority of cases with a recorded malignant site-specific diagnosis in their patient profile. This is consistent with results observed in **Chapter 5**, and with previously published results from Charlton *et al.*<sup>187</sup> and Haynes *et al.*<sup>188</sup>

##### **6.7.4.2 Linkage of primary care data to the cancer registry**

Linking of primary care data to external disease registries such as the NCDR is an evolving aspect of epidemiological studies that utilise electronic health records.

Inclusion of cancer events from either primary care or cancer registry data sources had little impact on the effect of statin use on cancer risk. In Chapter 5, PPV estimates of CPRD recorded cancer diagnoses were high across all cancers. In the other direction, sensitivity estimates of NCDR diagnosis varied by cancer type.

Based on the systematic review of identification of incident cancers in UK primary care databases (**Chapter 2**), a lower likelihood of confirmatory evidence would be found in the GP records of patients with a non-malignant diagnosis code. From the examination of NCDR linked data (**Chapter 5**); approximately 40% of cases identified in the NCDR alone (no concordant diagnosis in the CPRD) had a non-specific or non-malignant diagnosis code in the CPRD. In both circumstances, within

the statin-cancer association, the addition of these cases (either a patient from the probable/possible diagnostic group or a discordant case identified in the NCDR) had no substantial effect on study findings in the context of the statin-cancer association. In comparison to linked data relative risk estimates, estimates from the CPRD alone were higher for colorectal, lung, and prostate cancer, which may be due to an underestimation of cases in primary care as observed in **Chapter 6**. Primary care data linkage to cancer registry data would lead to a more precise estimate of overall cancer in comparison to utilising primary care data alone.

### **6.7.5 Residual bias in corrected analyses**

Confidence interval estimates from some “corrected” analyses did not cross 1.

“Corrected” analysis of immortal, protopathic and prevalent user bias produced confidence intervals  $>1$  for breast, colorectal, and lung cancer; an increased risk of prostate cancer was the most pronounced effect. Unmeasured confounding may explain why the corrected analysis showed a result that was in some cases further from the null (assumed true association) than the biased analysis. From the systematic review (**Chapter 3**), prostate cancer showed the most variation in terms of observed effect: 3 studies observed an increased risk, 4 observed a reduced risk, and 10 observed no association. Detection bias, particularly for prostate cancer, is the main argument given by previous studies that observed an increased risk of cancer among statin users.<sup>207</sup> This study adjusted for consultation rate and also for healthy user bias by considering the comparison group of glaucoma medication users. Although the null effect observed for most cancer types in the healthy user analysis is consistent with detection bias circumvented by employing an active

comparison drug group, the impact of the bias is uncertain because of the lower power and precision due to smaller number of patients overall.

In comparison to the cohort design analyses, corrected relative risk estimates from the nested case-control study were in range of what was assumed to be the true association between statin use and cancer risk. The variability between designs could have contributed to the differences observed between the corrected relative risk estimates from the cohort study designs (immortal, protopathic, prevalent user bias) compared to the nested case-control study design (time-window bias). As reported by Madigan *et al.*<sup>211</sup> variability between study designs has been shown to be a driving factor between studies examining the same question. In contrast, variability within study design had a lower impact on study findings which is consistent with observed findings from bias analysis of immortal, protopathic, and prevalent user bias, which employed the matched cohort design.

#### **6.7.6 Limitations**

This study had several limitations. First, results are limited to one particular type of drug (statins) and to cancers of the breast, colorectum, lung, and prostate. In addition, the UK CPRD stores primary care data collected from UK general practices. Findings may have differed if other drug-cancer pairings had been investigated, or if another data source had been utilised.

Second, the biases examined in this thesis are not exhaustive; for example, bias related to adjustment of unmeasured confounders, measurement error, or missing data were not examined. Although, in their own right, they are important design considerations to bear in mind when conducting an observational study,

examination of their impact in a pharmacoepidemiological setting was outside the scope of this thesis.

Third, the choice of study design was not exhaustive; all possible design options were not implemented. For example, varying the ratio of controls to cases for time-window bias, or changing the exposure definition to 1-year minimum exposure for immortal time bias would have been other possible options. The designs that were implemented in this study were intended to be representative of those used in past observational studies of the association between cancer and statin use (**Chapter 3**) or previous pharmacoepidemiological studies with cancer as the main outcome.

Lastly, the main objective of this study was not to estimate the statin-cancer association, but to estimate the impact of bias in the context of this question.

Adjustment for various potential confounders that could have affected the statin-cancer association was implemented. However, as with all observational studies, there was a chance of unmeasured confounding affecting results.

### **6.7.7 Future Studies**

Future studies may include further bias studies,<sup>212</sup> particularly for other drug-disease pairings. In addition, studies examining other biases, for example, those related to missing data, unmeasured confounding, or measurement error could be conducted in real settings i.e. empirically from an electronic healthcare database or alternatively by simulated data.

### **6.7.8 Conclusion**

Pharmacoepidemiological studies have the possibility to impact public health and influence the decision making process of both clinician and patient. A major concern, however, has been conflicting findings from recent

pharmacoepidemiological studies,<sup>19, 21</sup> which could put both clinician and patient in uncertain positions with regards to prescribing and either initiation or continuation of such medications, respectively.

A number of observational studies examining the risk of cancer associated with statin use have shown conflicting findings (**Chapter 3**). A number of common design flaws and decisions have been postulated as drivers of these discrepant results. However, this chapter has demonstrated that in a practical study setting, these flaws and differences in study design do not uniformly lead to large changes in estimated associations between statin use and cancer risk. Only time-window bias lead to consistent differences between biased and corrected analyses. In contrast, none of the other postulated biases or differing case definitions substantially influenced the perceived risk of cancer associated with statin use.

Nevertheless, study-specific factors are likely to affect the magnitude of different biases in different settings. Therefore, appropriate selection of design methods and sensitivity/bias analysis are needed to ensure transparency and confidence in study conclusions, particularly if results divert from past findings.

## **6.8 Summary**

- A series of observational studies were undertaken within the context of the statin-cancer association to measure and compare the impact of several potential drivers of conflicting findings including study bias, case definitions, and data linkage.
- Assessed biases included immortal time, protopathic, prevalent user, healthy user, and time-window bias.

- Of the seven potential drivers of discrepant results in the example study of statins and cancer, only time-window bias yielded substantial and consistent biased effects, with bias towards a protective association and corrected analyses yielding a null association.
- Immortal time, protopathic, prevalent user, and healthy user bias had minimal impact on the estimated association between statin use and cancer risk.

## 7 Thesis summary and conclusions

### 7.1 Introduction

In this chapter, the main findings and key discussion points that have been described in this thesis are summarised and brought together. First, a summary of the research undertaken is outlined. Second, for each main analysis (**Chapters 5 and 6**), key findings are compared with past research reviewed in **Chapters 2-3**, and their strengths and limitations summarised. Last, implications, areas of future research, and conclusions are outlined.

### 7.2 Summary of research undertaken

In recent years, an increasing number of pharmacoepidemiological studies have been undertaken using databases of routinely collected health records. However, there have been conflicting findings from studies examining the same question using similar databases.

The primary aim of this thesis was to examine whether differences in case ascertainment or common design flaws could explain conflicting findings among studies examining the association between statin use and cancer risk.

As a starting point, two systematic reviews of existing literature were undertaken.

The first examined current practices used to identify incident cancer from UK primary care databases (**Chapter 2**). The second evaluated findings from observational studies examining the risk of cancer associated with statin use, with a focus on methodological considerations (**Chapter 3**).

Two main analyses were conducted to address the main aim of this thesis. The first, a validation study of recorded cancer diagnoses in the CPRD (**Chapter 5**). The second, a series of observational studies using primary care data from the CPRD to

measure the impact of several study design flaws and decisions on the association between statin use and cancer risk (**Chapter 6**).

### **7.3 Validity of cancer diagnosis in the CPRD (Chapter 5)**

#### **7.3.1 Summary of main findings from Chapter 5**

- i.* Two case definitions were developed from primary care data to estimate cancer incidence rates:
  - a.* Standard case definition: malignant site-specific diagnoses.
  - b.* Broad case definition: inclusion of cases with borderline, suspected or general codes with supporting evidence of diagnosis (broad definition) were.
- ii.* In a random sample of 2 million patients from the CPRD, estimated incidence rates for breast, colorectal, lung, and prostate cancer were compared to national rates published by the ONS. In comparison to national rates, primary care incidence rates were lower across all cancer types examined. Disparities varied by age, sex, calendar year, and cancer type.
- iii.* Estimated primary care incidence rates of cancer were marginally increased by the inclusion of non-specific cancer diagnoses (broad case definition), but remained lower than nationally published rates.
- iv.* In an analysis of agreement between primary care data (CPRD) and linked cancer registry data (NCDR), in terms of capturing recorded cancer diagnoses the NCDR was treated as the gold standard. PPV estimates of CPRD recorded cancer diagnoses were generally high, ranging from 88% for prostate cancer to 90% for colorectal cancer.





- vii.** Higher (in comparison to ONS published rates) linked incidence rates may be partly explained by possible false-positive cases identified in primary care data, or cases that were not registered nationally. However, future studies need to confirm or refute this finding.

### **7.3.2 Validity of recorded cancer diagnoses in the context of previous research**

- i.** Consistent with previous studies,<sup>187, 188</sup> estimated cancer incidence rates from primary care data were lower compared to national rates.
- ii.** Consistent with Charlton *et al.*<sup>187</sup> the addition of colorectal cancer cases with non-specific cancer related codes resulted in marginal increases of estimated incidence rates. In this study, similar findings were also observed for breast, lung, and prostate cancer.
- iii.** Consistent with previous studies, high estimates of positive predictive values of recorded primary care diagnosis with respect to recorded cancer registrations were observed. However, a small proportion of cases in the CPRD did not have a match in the cancer registry. Whether these cases were not registered nationally or are false-positive cases is uncertain.
- iv.** In contrast to past cancer registry linkage studies,<sup>80, 194</sup> sensitivity of recorded primary care cancer diagnoses with respect to the cancer registry in this study was low. Possible reasons why this may be the case include:
  - a.** Agreement was defined as specific matches of ICD-10 (breast: C50; colorectum: C18-C20; lung: C34; and prostate: C61) codes between all data sources. Whether previous studies utilised broader definitions is unclear from reported methods.

- b.* Cohorts with pre-existing conditions (diabetes,<sup>194</sup> symptoms of cancer<sup>80</sup> were used in previous studies, this might have overestimated agreement as these patients may be more likely to have cancer detected.
- c.* A different version of the linked dataset was used. This study utilised Set 9, while previous studies may have used earlier versions.
- d.* This study did not access linked hospital episodes statistics data or free-text, which may have increased sensitivity estimates if utilised.
- v.* No other studies have estimated cancer incidence rates from linked primary care data. However, a study linking primary care data (CPRD) to the national registry of acute coronary syndromes (Myocardial Ischaemia National Audit Project, MINAP) also observed higher incidence rates of acute myocardial infarction compared to expected rates when using combined data from the CPRD, MINAP, Hospital Episodes Statistics, and ONS mortality.<sup>196</sup>

### **7.3.3 Strengths of this study**

This validation study has several strengths:

- i.* In this thesis, a random sample of the CPRD was used to evaluate agreement between linked data sources. In contrast, previous studies have focussed on specific clinical groups, e.g. diabetes patients,<sup>194</sup> with limited generalisability.
- ii.* This study used data from a 10-year period, and was one of the largest studies to examine cancer incidence rate estimates using primary care data.
- iii.* The cancer outcome code list modified for purposes of this study included a broad range of codes for most cancer types. Both malignant and non-

malignant cancer related diagnosis codes were included. The broad coverage of the code list enabled comprehensive identification of cases as well as identification of cancer related diagnoses from patient profiles in the CPRD or NCDR when discordance occurred.

#### **7.3.4 Limitations of the study**

Limitations of this study are detailed in **Chapter 5**. A brief summary of these

limitations are listed below:

- i.* Findings from this study were limited to cancers of the breast, colorectum, lung, and prostate and may not be generalised to other cancer types.
- ii.* Linkage of primary care data to external sources were limited to participating patients and practices and may not be generalised to the CPRD as a whole. However, incidence rate differences were marginal when comparing estimates from the initial CPRD random sample to estimates from linkage eligible patients and practices.
- iii.* In this study, agreement of recorded diagnoses was defined as a specific match of ICD-10 codes in all data sources (CPRD, NCDR, and ONS mortality), which may be considered narrow. For example, a Read code indicating malignant cancer of the colon (e.g. B13..00) would be mapped to C18 (C18: Malignant neoplasm of the colon). If this particular diagnosis was coded in the NCDR with a related code such as C26 (Malignant neoplasm of ill-defined digestive organ), then this would be considered a discordant match. However, for all discordant cases, searches of other relevant recorded diagnoses in both the CPRD and NCDR were examined in an attempt to explain the disagreement between data sources.

## 7.4 Systematic evaluation of the impact of potential methodological drivers of discrepant results in a pharmacoepidemiological study of statin use and cancer risk (Chapter 6)

### 7.4.1 Summary of main findings from Chapter 6

- i.* In a cohort of *new statin* users matched to non-users, the impact of immortal time bias in the context of the statin-cancer association was marginal. There was a consistent direction of effect closer to the null, biased analyses resulted in lower relative risk estimates compared to corrected analyses.  $\Delta\beta$  was consistently estimated at 0.01 across all cancer types examined.
- ii.* Also in a new user matched cohort, the effect of protopathic bias was minimal.  $\Delta\beta$  estimates for protopathic bias ranged from 0.00 for colorectal cancer to 0.03 for breast cancer.
- iii.* In an analysis to assess the impact of prevalent user bias, a cohort consisting of both prevalent and *new statin* users was compared to a new user cohort of statin users. Inclusion of prevalent users moved the effect estimate toward the null suggesting a weak protective or null association from the inclusion of prevalent users. The impact of prevalent user bias was minimal,  $\Delta\beta$  estimates for prevalent user bias ranged from -0.09 for breast cancer to -0.05 for prostate cancer.
- iv.* The impact of healthy user bias was assessed by comparing two cohorts. Cohort 1 (corrected analysis) included a comparison of *new statin* users and new users of glaucoma medication users. The second cohort (potentially biased analysis) comprised of statin users matched to non-users. Effects of

healthy user bias varied by cancer type:  $\Delta\beta$  ranged from -0.05 (95% CI; -0.18, 0.09) for colorectal cancer to 0.07 (95% CI; -0.06, 0.19) for breast cancer.

- v.** Time-window bias yielded the largest impact among the different biases considered. For all cancer types, there was a consistent direction of effect from biased (protective association) to corrected analysis (null association).  $\Delta\beta$  estimates between biased (time-independent sampling of controls) and corrected analyses (risk-set sampling of controls) ranged from -0.34 (95% CI; -0.37, -0.25) for lung cancer to -0.13 (95% CI; -0.19, -0.08) for prostate cancer.
- vi.** From the systematic review described in **Chapter 2** two case definitions were typically used to define cancer outcome in primary care data. In comparison to the standard case definition which required one definite malignant diagnosis code, differences in relative risk estimates from the broad case definition (addition cases with non-specific cancer related diagnosis codes) were marginal.  $\Delta\beta$  ranged from -0.01 for breast and colorectal cancer to -0.03 for lung and prostate cancer.
- vii.** Overall, the incorporation of patient-level linked data from both the cancer and death registry generated similar results to those that used only primary care data restricted to linkage eligible patients and practices.  $\Delta\beta$  due to linked data (CPRD *OR* linked data) ranged from 0.01 (95% CI; -0.09, 0.11) for prostate cancer to 0.05 (95% CI; -0.06, 0.16) for colorectal cancer.

#### **7.4.2 Impact of potential drivers of conflicting results in the context of previous research**

- i.** There was generally varied methodology among 30 studies investigating the association between statin use and risk of breast, colorectal, and prostate

cancer (**Chapter 3**). Furthermore, several populations and sources of electronic data were utilised by these studies including routinely collected health data from GP surgeries, administrative data from insurance claims, and disease registries. Variables collected by these sources as well as reasons for data collection vary between these data sources. Comparisons of findings from the reviewed studies and the study presented in this thesis should be made cautiously.

- ii.* From the 30 studies, there were several studies reporting a statistically significant increased or reduced risk of cancer, particularly for colorectal, lung and prostate cancer. Some evidence of consistent effects from prevalent user and time-window bias was observed. Correction for healthy user bias yielded the most variability in terms of relative risk estimates.
- iii.* Many of the examples of immortal time bias described in literature have shown relatively large differences in risk estimates between biased and corrected analyses.<sup>69, 213</sup> In contrast, findings from this study have shown minimal impact of immortal time bias. Two possible reasons could explain this conflict:
  - a.* In previous research, immortal time bias had not been examined in the context of the statin cancer association. Typical definitions of exposure status from studies may depend on the drug disease pairing (e.g acute diseases) investigated which may influence the magnitude of immortal time bias.
  - b.* Extreme design parameters may have been implemented in these examples leading to a greater impact of immortal time bias.

- iv.** In the context of the systematic review conducted in **Chapter 3**, 1 out of 16 cohort studies may have potentially been affected by immortal time bias. The majority of studies in the review defined exposure status as  $\geq 1$  statin prescription. Based on these definitions, the potential impact of immortal time bias is likely to be minimal due to the direct correlation between exposure definitions and magnitude of immortal time.
- v.** From the systematic review conducted in **Chapter 3**, 12 studies did not implement a lag-period to guard from protopathic bias. There were suggestions of this bias influencing findings, particularly when the 4/12 studies linked short term statin use ( $\leq 12$  months) with increased/decreased risk of cancer. However, these findings were difficult to un-tangle as many of these studies may have also been affected by prevalent user bias which may have jointly influenced spurious findings of short term associations.
- vi.** In this study, the impact of prevalent user bias was minimal. However, weak evidence of a protective or null association from the inclusion of prevalent users was observed. This effect was consistent with studies reviewed in **Chapter 3**: two new user cohort studies, Smeeth *et al.* and Hippisley-Cox *et al.* presented higher relative risk estimates compared to their counterpart studies that included prevalent statin users.
- vii.** The effects of healthy user bias in this study were minimal. An active comparison group consisting of glaucoma medication users (prevent progression of glaucoma) was utilised to minimise differences in risk due to attitudes toward disease prevention and healthcare utilisation. Although the comparison groups were similar in some respects such as rate of GP



consultations prior to treatment initiation, there were differences observed in terms of prior co-medications and co-morbidities.

- viii.** 10 of 30 studies did not appear to account for healthy user bias (**Chapter 3**). None of 10 studies (i) used an active drug comparator group; or (ii) adjusted for health utilisation services such as screening or cancer related testing. Among the 30 review studies, attempts to minimise healthy user bias by either by (i) or (ii) yielded the most variability in study conclusions compared to other biases considered.
- ix.** The current study was in line with reported findings from Setoguchi *et al.*<sup>146</sup> (null association); in contrast, Clancy *et al.*<sup>214</sup> reported a decreased risk of colorectal cancer when comparing new statin users to new glaucoma medication users.
- x.** Time-window bias yielded the largest impact of all biases considered in this thesis. However, only one of the reviewed case-control studies was susceptible to this bias (**Chapter 3**). Consistent with findings from this study, the impact of this bias has been shown in case-control studies examining lung and pancreatic cancer risk associated with statin use.<sup>119 28</sup>
- xi.** No other studies have evaluated the effect of alternative cancer case definitions on epidemiological study findings in UK primary care databases.
- xii.** No other studies have looked at various forms of linked data in a pharmacoepidemiological setting. However, Boggon *et al.*<sup>88</sup> compared cancer survival from the CPRD to CPRD linked cancer registry data, and found marginal difference in cancer survival estimates.

*xiii.* Importantly, absolute risk estimates of statin use and cancer risk were higher when utilising primary care data alone compared to estimates incorporating linked cancer registry data. In line with results from **Chapter 5**, cancer risk may be underestimated if only primary care data is utilised.

#### **7.4.3 Strengths of the study**

- i.* In comparison to previous studies of statin use and cancer risk in UK primary care databases,<sup>56, 128, 130, 131</sup> this study had similar if not greater power.
- ii.* Prescription data in the CPRD are automatically captured when GPs issue prescriptions. Therefore complete records of prescriptions written by the GP are stored on the CPRD which provides assurance about the quality of the data going into the varying definitions of exposure implemented in this thesis. However, there is no information on whether medications had actually been taken or adhered to the prescribed course. That being said, if repeat prescriptions had been written by the GP over several occasions the patient is likely to have been taking these medications.
- iii.* Design parameters (such as exposure definition, outcome definitions) reflected those implemented by past studies described in **Chapter 2 and 3**.
- iv.* Findings from this study appear to be robust. Several sensitivity analyses were conducted to assess the effect of missing data and matching with replacement. In both cases, similar  $\Delta\beta$  estimates to those from the primary analyses were observed.

#### **7.4.4 Limitations of the study**

- i.* Cancers of the breast, colorectum, lung, and prostate were examined in this study. Therefore, findings from this study may not be generalised to other cancer types.
- ii.* This study was conducted in the UK CPRD and results may not be generalised to studies settings outside the UK.
- iii.* Unlike a simulated data setting, not all design decisions could be controlled (e.g. unmeasured confounding). However, intricacies of “real” data such as recorded prescription data or variability in GP recorded diagnoses codes may be difficult to mimic in simulated data.
- iv.* Potential for unmeasured confounding is a possibility in all observational studies. However, the main aim of this study was to estimate the impact of potential drivers of discrepant results and not the true association between statin use and cancer risk.

#### **7.4.5 Bias analyses considerations for application to other exemplars of pharmacoepidemiological research**

While the present study has led to some clear conclusions regarding drivers of bias in statin-cancer studies, the same lessons will not all necessarily carry over to other drug-outcome association studies. First, studies of drugs other than statins are likely to have different issues: statins are very commonly prescribed, used for long-term prevention, indications may differ to other drugs, and intended to be taken for life. Other drugs may be less prevalent, used for acute conditions, or prescribed for the short term. Second, this study focussed on cancer outcomes, which are uncommon, latent diseases that may take some time to present. Other outcomes are likely to have different issues such as acute events including fracture risk or

myocardial infarction, which may occur directly after exposure, or soon after. Last, different data sources record patient data for different reasons; for example, smoking status is captured in the CPRD, in contrast, US claims databases may not necessarily record lifestyle factors such as smoking status, alcohol status, or BMI. Non-availability or non-inclusion of important confounders may lead to residual confounding and limit the interpretability of a study. Alternatively, simulation of drug-disease associations could act as a secondary or main analysis to circumvent residual confounding by assuming all confounding factors have been measured and included in the statistical model.

Although both immortal time and protopathic bias had minimal impact on the statin-cancer association in this thesis, examples of their impact have been well noted in other drug-disease associations. For example, protopathic bias has been shown to drive an increased risk of gastric cancer among patients prescribed proton pump inhibitors (PPIs).<sup>116</sup> Patients may be prescribed PPIs to address symptoms of undetected gastric cancer such as gastric ulcers, and in turn, use of PPIs may be incorrectly attributed to the cause of the gastric cancer once detected. In contrast, indications such as lipid levels, which may initiate statin therapy, may not be directly related to early symptoms of the cancer types examined in this thesis. Immortal time bias also had minimal effect on the statin-cancer association in this thesis. However, acute outcomes such as myocardial infarction or fracture risk may occur shortly after exposure which may influence immortal time bias and drive study conclusions.<sup>69</sup>

In spite of likely varying issues for different drug-disease studies, the overall broad approach to investigating bias was useful and could be carried over to examine the impact of methodological variation in other pharmacoepidemiology contexts.

## **7.5 Implications of research undertaken**

In this thesis and previous studies, lower rates of recorded diagnoses have been observed for breast, colorectal, lung, and prostate cancer in primary care in comparison to national rates. However, linkage to the cancer registry has shown that some of these cases have been registered nationally, but not recorded in primary care data, particularly for the elderly and for cancer types with high mortality. Recording of events (e.g. clinical diagnoses, referrals) by GPs is mainly used for day-to-day management of patients and not for research purposes. However, an implication of this observed disparity between primary care and national estimates is the need for linked cancer registry data when complete ascertainment of events is important; a potential example may include population based studies investigating healthcare interventions. In the other direction, there were a small minority of cases identified in primary care data, but not in cancer registry data. There is a possibility that these cases are being excluded from national registries due to inconclusive evidence of diagnosis; and case status in primary care not subsequently updated. This may call for better processes in consolidating records between data sources. However, this would be difficult with regard to the Read coding dictionary, as a code to refute or un-confirm previous diagnoses is not currently available, and would call for an update to the Read coding dictionary to enable recording of this process.

Recent years have seen a surge of methodological studies attempting to explain conflicting findings from studies utilising routine collected health data.<sup>69, 119, 208, 211, 212, 215</sup> From this study, most design flaws implemented did not show substantial impact on the statin cancer association. However, two implications for future pharmacoepidemiological studies arise based on findings from this study. First, careful consideration of how design choices might affect study results is needed e.g. choice of database or control sampling. Second, if study results have deviated from previous findings then relevant sensitivity analyses should be conducted to ensure robust findings.

The empirical case-study conducted as part of this thesis

## **7.6 Future research**

Part of this thesis utilised linked cancer registry data to assess the validity of recorded diagnoses by comparing linked primary care incidence rates to that of expected rates based on cancer registrations published by the ONS. Cancers of the breast, colorectum, lung, and prostate were examined; however, future studies could consider other cancer types. Moreover, future studies could confirm or refute the apparently higher incidence of cancer types investigated when incorporating linked cancer registry data as well as whether this increase occurs for other cancer types. In addition, the completeness and utilisation of additional cancer registry variables such as treatment and disease severity variables (e.g. stage and grade) could be used to add further dimensions to observational drug safety studies.

In terms of conflicting findings between studies, part of the variability can be explained by design flaws examined in this thesis. However, further studies could examine alternative biases (e.g. measurement error, missing data) in settings which

use “real” or simulated data. With regards to the latter, a wider range of situations and design decisions could be explored which could help to identify specific situations that might lead to a larger impact. Additionally, other drug disease pairings which have been of concern (e.g. diabetes and the risk of cancer) could also be examined in a bias study setting. Incidence of disease, utilisation of health services, and prevalence of prescribed medications vary between populations. Different populations may have contributed to part of the variation in conflicting findings between studies.<sup>215</sup> A possible solution could be a meta-analysis or separate analyses by data source (population) based on a standardised protocol, then assessing whether findings are consistent between populations.<sup>215</sup>

## **7.7 Conclusions**

In line with previous studies,<sup>196</sup> findings from this study have shown that sole use of primary care data to identify particular cancer outcomes may be biased and lead to an underestimation of cancer incidence. Primary care data may misclassify case status without linkage to external data sources such as the cancer or death registry, particularly among elderly patients, and for cancer types that are captured at later stages of disease progression with high mortality. Failure to incorporate linked data from the cancer and death registry may result in the exclusion of false-negative cases that have not been identified in primary care data. On the other hand, utilisation of linked data may produce higher incidence of cancers either due to false-positive cases in primary care data or the possibility that they are simply not registered nationally.

The effect of bias can influence results, and sway findings in either direction.

However, their effect can also be minimal as shown in this thesis. Observational

studies are inherently prone to bias, and therefore the potential impact of such biases should be evaluated to ensure confidence and transparency in findings. This was once a difficult task for several reasons including: lack of computing power; difficulty in sharing large datasets; time-constraints; and the labour intensive nature of conducting an analysis. However, electronic health databases have made various sensitivity analyses and even re-analysis of studies a feasible option.<sup>19</sup> Re-analysis using different methodologies could rule out or identify possible factors that could drive conflicting results between studies.



## References

1. Griffin JP. *The textbook of pharmaceutical medicine*. 6th ed. Oxford: Wiley-Blackwell, 2009.
2. Hawton K, Bergen H, Simkin S, *et al*. Effect of withdrawal of co-proxamol on prescribing and deaths from drug poisoning in England and Wales: time series analysis. *BMJ* 2009;338:b2270.
3. Yellow Card Scheme. from <https://yellowcard.mhra.gov.uk/>.
4. McNaughton R, Huet G, Shakir S. An investigation into drug products withdrawn from the EU market between 2002 and 2011 for safety reasons and the evidence used to support the decision-making. *BMJ open* 2014;4(1):e004221.
5. Analytical Trend Troubles Scientists. from <http://www.wsj.com/articles/SB10001424052702303916904577377841427001840>.
6. Rothman KJ, Greenland S, Lash TL. *Modern epidemiology*. 3rd ed. Philadelphia ; London: Lippincott Williams & Wilkins, 2008.
7. Chen YC, Wu JC, Haschler I, *et al*. Academic impact of a public electronic health database: bibliometric analysis of studies using the general practice research database. *PloS one* 2011;6(6):e21404.
8. Weiner J. A comparison of primary care systems in the USA, Denmark, Finland and Sweden: lessons for Scandinavia? *Scandinavian journal of primary health care* 1988;6(1):13-27.
9. Kaiser Permanente Research. from <http://www.dor.kaiser.org/external/research/topics/Pharmacoepidemiology/>.
10. Hines DM, McGuinness CB, Schlienger RG, *et al*. Incidence of ischemic colitis in treated, commercially insured hypertensive adults: a cohort study of US health claims data. *American journal of cardiovascular drugs : drugs, devices, and other interventions* 2015;15(2):135-49.
11. Schuck-Paim C, Taylor R, Lindley D, *et al*. Use of near-real-time medical claims data to generate timely vaccine coverage estimates in the US: the dynamics of PCV13 vaccine uptake. *Vaccine* 2013;31(50):5983-8.
12. Petri H, Maldonato D, Robinson NJ. Data-driven identification of co-morbidities associated with rheumatoid arthritis in a large US health plan claims database. *BMC musculoskeletal disorders* 2010;11:247.

13. Pelletier EM, Shim B, Ben-Joseph R, *et al.* Economic outcomes associated with microvascular complications of type 2 diabetes mellitus: results from a US claims data analysis. *Pharmacoeconomics* 2009;27(6):479-90.
14. Pelletier EM, Shim B, Goodman S, *et al.* Epidemiology and economic burden of brain metastases among patients with primary breast cancer: results from a US claims data analysis. *Breast cancer research and treatment* 2008;108(2):297-305.
15. Robinson D, Jr., Hackett M, Wong J, *et al.* Co-occurrence and comorbidities in patients with immune-mediated inflammatory disorders: an exploration using US healthcare claims data, 2001-2002. *Current medical research and opinion* 2006;22(5):989-1000.
16. Furu K, Wettermark B, Andersen M, *et al.* The Nordic countries as a cohort for pharmacoepidemiological research. *Basic & clinical pharmacology & toxicology* 2010;106(2):86-94.
17. Clinical Practice Research Datalink Bibliography (internet). Retrieved 10 April, 2013, from [www.cprd.com/bibliography/](http://www.cprd.com/bibliography/).
18. Boston Collaborative Drug Surveillance Program. Retrieved 10 April, 2013, from <http://www.bu.edu/bcdsp/publications-2/>.
19. de Vries F, de Vries C, Cooper C, *et al.* Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the General Practice Research Database. *International journal of epidemiology* 2006;35(5):1301-8.
20. Meier CR, Schlienger RG, Kraenzlin ME, *et al.* Statin drugs and the risk of fracture. *Jama* 2000;284(15):1921-2.
21. Dixon WG and Solomon DH. Bisphosphonates and esophageal cancer--a pathway through the confusion. *Nature reviews. Rheumatology* 2011;7(6):369-72.
22. Green J, Czanner G, Reeves G, *et al.* Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. *BMJ* 2010;341:c4444.
23. Cardwell CR, Abnet CC, Cantwell MM, *et al.* Exposure to oral bisphosphonates and risk of esophageal cancer. *Jama* 2010;304(6):657-63.
24. Chang CC, Ho SC, Chiu HF, *et al.* Statins increase the risk of prostate cancer: a population-based case-control study. *The Prostate* 2011;71(16):1818-24.

25. Farwell WR, D'Avolio LW, Scranton RE, *et al.* Statins and prostate cancer diagnosis and grade in a veterans population. *Journal of the National Cancer Institute* 2011;103(11):885-92.
26. Farwell WR, Scranton RE, Lawler EV, *et al.* The association between statins and cancer incidence in a veterans population. *Journal of the National Cancer Institute* 2008;100(2):134-9.
27. Khurana V, Bejjanki HR, Caldito G, *et al.* Statins reduce the risk of lung cancer in humans: a large case-control study of US veterans. *Chest* 2007;131(5):1282-8.
28. Khurana V, Sheth A, Caldito G, *et al.* Statins reduce the risk of pancreatic cancer in humans: a case-control study of half a million veterans. *Pancreas* 2007;34(2):260-5.
29. Murtola TJ, Tammela TL, Maattanen L, *et al.* Prostate cancer and PSA among statin users in the Finnish prostate cancer screening trial. *International journal of cancer. Journal international du cancer* 2010;127(7):1650-9.
30. Smeeth L, Douglas I, Hall AJ, *et al.* Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. *British journal of clinical pharmacology* 2009;67(1):99-109.
31. Meier CR, Schlienger RG, Kraenzlin ME, *et al.* HMG-CoA reductase inhibitors and the risk of fractures. *Jama* 2000;283(24):3205-10.
32. Hemkens LG, Bender R, Grouven U, *et al.* Insulin glargine and cancer. *Lancet* 2009;374(9703):1743-4; author reply 44.
33. Currie CJ, Poole CD, Gale EA. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. *Diabetologia* 2009;52(9):1766-77.
34. Pocock SJ and Smeeth L. Insulin glargine and malignancy: an unwarranted alarm. *Lancet* 2009;374(9689):511-3.
35. Suissa S and Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes care* 2012;35(12):2665-73.
36. Walley T, Folino-Gallo P, Stephens P, *et al.* Trends in prescribing and utilization of statins and other lipid lowering drugs across Europe 1997-2003. *British journal of clinical pharmacology* 2005;60(5):543-51.
37. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 1994;344(8934):1383-9.

38. Shepherd J, Cobbe SM, Ford I, *et al.* Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. 1995. *Atherosclerosis. Supplements* 2004;5(3):91-7.
39. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. *The New England journal of medicine* 1998;339(19):1349-57.
40. Downs JR, Clearfield M, Weis S, *et al.* Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. Air Force/Texas Coronary Atherosclerosis Prevention Study. *Jama* 1998;279(20):1615-22.
41. Shepherd J, Blauw GJ, Murphy MB, *et al.* Pravastatin in elderly individuals at risk of vascular disease (PROSPER): a randomised controlled trial. *Lancet* 2002;360(9346):1623-30.
42. Baron JA. Statins and the colorectum: hope for chemoprevention? *Cancer Prev Res (Phila)* 2010;3(5):573-5.
43. Chan KK, Oza AM, Siu LL. The statins as anticancer agents. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2003;9(1):10-9.
44. Ciofu C. The statins as anticancer agents. *Maedica* 2012;7(4):377.
45. Etminan M, Gill S, Samii A. The role of lipid-lowering drugs in cognitive function: a meta-analysis of observational studies. *Pharmacotherapy* 2003;23(6):726-30.
46. Stewart BW and Wild C. *World Cancer Report 2014.*
47. Tyczynski JE, Bray F, Parkin DM. Lung cancer in Europe in 2000: epidemiology, prevention, and early detection. *The Lancet. Oncology* 2003;4(1):45-55.
48. Newman TB and Hulley SB. Carcinogenicity of lipid-lowering drugs. *JAMA* 1996;275(1):55-60.
49. Sacks FM, Pfeffer MA, Moyer LA, *et al.* The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. Cholesterol and Recurrent Events Trial investigators. *The New England journal of medicine* 1996;335(14):1001-9.

50. West of Scotland Coronary Prevention Study: identification of high-risk groups and comparison with other cardiovascular intervention trials. *Lancet* 1996;348(9038):1339-42.
51. Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT-LLT). *Jama* 2002;288(23):2998-3007.
52. Bonovas S, Filioussi K, Tsavaris N, *et al.* Statins and cancer risk: a literature-based meta-analysis and meta-regression analysis of 35 randomized controlled trials. *J Clin Oncol* 2006;24(30):4808-17.
53. Bonovas S, Filioussi K, Tsavaris N, *et al.* Use of statins and breast cancer: a meta-analysis of seven randomized clinical trials and nine observational studies. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2005;23(34):8606-12.
54. Cauley JA, McTiernan A, Rodabough RJ, *et al.* Statin use and breast cancer: prospective results from the Women's Health Initiative. *Journal of the National Cancer Institute* 2006;98(10):700-7.
55. Coogan PF, Rosenberg L, Palmer JR, *et al.* Statin use and the risk of breast and prostate cancer. *Epidemiology* 2002;13(3):262-7.
56. Smeeth L, Douglas I, Hall AJ, *et al.* Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. *British Journal of Clinical Pharmacology* 2009;67(1):99-109.
57. Bonovas S, Filioussi K, Flordellis CS, *et al.* Statins and the risk of colorectal cancer: a meta-analysis of 18 studies involving more than 1.5 million patients. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2007;25(23):3462-8.
58. Poynter JN, Gruber SB, Higgins PD, *et al.* Statins and the risk of colorectal cancer. *The New England journal of medicine* 2005;352(21):2184-92.
59. Blais L, Desgagne A, LeLorier J. 3-Hydroxy-3-methylglutaryl coenzyme A reductase inhibitors and the risk of cancer: a nested case-control study. *Archives of internal medicine* 2000;160(15):2363-8.
60. Graaf MR, Beiderbeck AB, Egberts AC, *et al.* The risk of cancer in users of statins. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2004;22(12):2388-94.

61. Coogan PF, Rosenberg L, Strom BL. Statin use and the risk of 10 cancers. *Epidemiology* 2007;18(2):213-9.
62. Kaye JA and Jick H. Statin use and cancer risk in the General Practice Research Database. *British journal of cancer* 2004;90(3):635-7.
63. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to statins and risk of common cancers: a series of nested case-control studies. *BMC cancer* 2011;11:409.
64. Singh H, Mahmud SM, Turner D, *et al.* Long-term use of statins and risk of colorectal cancer: a population-based study. *American Journal of Gastroenterology* 2009;104(12):3015-23.
65. Shannon J, Tewoderos S, Garzotto M, *et al.* Statins and prostate cancer risk: a case-control study. *American journal of epidemiology* 2005;162(4):318-25.
66. Statins and prostate cancer risk: a large case-control study in veterans. ASCO Annual Meeting Proceedings 2005. *J Clin Oncol.*
67. Haukka J, Sankila R, Klaukka T, *et al.* Incidence of cancer and statin usage--record linkage study. *International journal of cancer. Journal international du cancer* 2010;126(1):279-84.
68. Chang CC, Ho SC, Chiu HF, *et al.* Statins increase the risk of prostate cancer: A population-based case-control study. *Prostate* 2011;71(16):1818-24.
69. Suissa S. Immortal time bias in pharmaco-epidemiology. *American journal of epidemiology* 2008;167(4):492-9.
70. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *American journal of epidemiology* 2003;158(9):915-20.
71. Horwitz RI and Feinstein AR. The problem of "protopathic bias" in case-control studies. *The American journal of medicine* 1980;68(2):255-8.
72. The Health Improvement Network Bibliography (internet). Retrieved 10 April, 2013, from <http://csdmruk.cegedim.com/THINBibliography.pdf>.
73. QRESEARCH Bibliography (internet). Retrieved 10 April, 2013, from <http://www.qresearch.org/SitePages/publications.aspx>.
74. General Practice Notebook - a UK medical reference (internet). Retrieved 1 March, 2013, from <http://www.gpnotebook.co.uk/>.
75. Patient.co.uk (internet). Retrieved 10 April, 2013, from [www.patient.co.uk](http://www.patient.co.uk).

76. *International classification of diseases for oncology : ICD-O / editors, April Fritz ... [et al.]*. Geneva :: World Health Organization, 2000.
77. Emberson JR, Ng LL, Armitage J, *et al*. N-terminal Pro-B-type natriuretic peptide, vascular disease risk, and cholesterol reduction among 20,536 patients in the MRC/BHF heart protection study. *Journal of the American College of Cardiology* 2007;49(3):311-9.
78. Collins R, Armitage J, Parish S, *et al*. MRC/BHF Heart Protection Study of cholesterol-lowering with simvastatin in 5963 people with diabetes: a randomised placebo-controlled trial. *Lancet* 2003;361(9374):2005-16.
79. World Health O. *ICD-10 : international statistical classification of diseases and related health problems / World Health Organization*. Geneva :: World Health Organization, 2004.
80. Dregan A, Moller H, Murray-Thomas T, *et al*. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer epidemiology* 2012;36(5):425-9.
81. Mackenzie IS, Macdonald TM, Thompson A, *et al*. Spironolactone and risk of incident breast cancer in women older than 55 years: retrospective, matched cohort study. *BMJ* 2012;345:e4447.
82. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to cyclooxygenase-2 inhibitors and risk of cancer: Nested case-control studies. *British journal of cancer* 2011;105(3):452-59.
83. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to bisphosphonates and risk of cancer: a protocol for nested case-control studies using the QResearch primary care database. *BMJ open* 2012;2(1):e000548.
84. Charlton R, Snowball J, Bloomfield K, *et al*. Colorectal cancer incidence on the General Practice Research Database. *Pharmacoepidemiology and drug safety* 2012.
85. Kaye JA, Derby LE, del Mar Melero-Montes M, *et al*. The incidence of breast cancer in the General Practice Research Database compared with national cancer registration data. *British journal of cancer* 2000;83(11):1556-8.
86. Gonzalez-Perez A and Garcia Rodriguez LA. Breast cancer risk among users of antidepressant medications. *Epidemiology* 2005;16(1):101-05.
87. Bodmer M, Becker C, Meier C, *et al*. Use of metformin is not associated with a decreased risk of colorectal cancer: a case-control analysis. *Cancer Epidemiology, Biomarkers & Prevention* 2012;21(2):280-6.

88. Boggon R, van Staa TP, Chapman M, *et al.* Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiology and drug safety* 2012.
89. Haynes K, Forde KA, Schinnar R, *et al.* Cancer incidence in the Health Improvement Network. *Pharmacoepidemiology and Drug Safety* 2009;18(8):730-36.
90. Vinogradova Y, Hippisley-Cox J, Coupland C, *et al.* Risk of Colorectal Cancer in Patients Prescribed Statins, Nonsteroidal Anti-Inflammatory Drugs, and Cyclooxygenase-2 Inhibitors: Nested Case-Control Study. *Gastroenterology* 2007;133(2):393-402.
91. Garcia Rodriguez LA and Huerta-Alvarez C. Reduced risk of colorectal cancer among long-term users of aspirin and nonaspirin nonsteroidal antiinflammatory drugs. *Epidemiology* 2001;12(1):88-93.
92. National Cancer Intelligence Network Cancer incidence by deprivation England, 1995-2004. (PDF 1.04MB) 2008.
93. Garcia Rodriguez LA and Gonzalez-Perez A. Risk of breast cancer among users of aspirin and other anti-inflammatory drugs. *British journal of cancer* 2004;91(3):525-29.
94. Garcia-Rodriguez LA and Huerta-Alvarez C. Reduced risk of colorectal cancer among long-term users of aspirin and nonaspirin nonsteroidal antiinflammatory drugs. *Epidemiology* 2001;12(1):88-93.
95. Haynes K, Forde KA, Schinnar R, *et al.* Cancer incidence in The Health Improvement Network. *Pharmacoepidemiology & Drug Safety* 2009;18(8):730-6.
96. Ronquist G, Rodriguez LAG, Ruigomez A, *et al.* Association between captopril, other antihypertensive drugs and risk of prostate cancer. *The Prostate* 2004;58(1):50-6.
97. Bhaskaran K, Douglas I, Evans S, *et al.* Angiotensin receptor blockers and risk of cancer: cohort study among people receiving antihypertensive drugs in UK General Practice Research Database. *Bmj* 2012;344:e2697.
98. Campbell J DJ, Eaton SC. Is the GPRD GOLD population comparable to the UK population? *29th International Conference on Pharmacoepidemiology & Therapeutic Risk Management; Montreal: Pharmacoepidemiol Drug Saf.* Montreal, Canada: Pharmacoepidemiology Drug Safety, 2013:280.
99. Dixon WG and Solomon DH. Bisphosphonates and esophageal cancer--a pathway through the confusion. *Nature Reviews Rheumatology* 2011;7(6):369-72.



100. Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. *Therapeutic Advances in Drug Safety* 2013;2042098613477445.
101. Benchimol EI, Langan S, Guttman A, *et al.* Call to RECORD: the need for complete reporting of research using routinely collected health data. *Journal of clinical epidemiology* 2013;66(7):703-5.
102. Bhaskaran K, Douglas I, Forbes H, *et al.* Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. *Lancet* 2014;384(9945):755-65.
103. Hippisley-Cox J and Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ open* 2015;5(3):e007825.
104. Couraud S, Dell'Aniello S, Bouganim N, *et al.* Cardiac glycosides and the risk of breast cancer in women with chronic heart failure and supraventricular arrhythmia. *Breast cancer research and treatment* 2014;146(3):619-26.
105. Jordan KP, Hayward RA, Blagojevic-Bucknall M, *et al.* Incidence of prostate, breast, lung and colorectal cancer following new consultation for musculoskeletal pain: a cohort study among UK primary care patients. *International journal of cancer. Journal international du cancer* 2013;133(3):713-20.
106. Muller S, Hider SL, Belcher J, *et al.* Is cancer associated with polymyalgia rheumatica? A cohort study in the General Practice Research Database. *Annals of the rheumatic diseases* 2014;73(10):1769-73.
107. Tsilidis KK, Capothanassi D, Allen NE, *et al.* Metformin does not affect cancer risk: a cohort study in the U.K. Clinical Practice Research Datalink analyzed like an intention-to-treat trial. *Diabetes care* 2014;37(9):2522-32.
108. Assayag J, Yin H, Benayoun S, *et al.* Androgen deprivation therapy and the risk of colorectal cancer in patients with prostate cancer. *Cancer causes & control : CCC* 2013;24(5):839-45.
109. Charlton RA, Snowball JM, Bloomfield K, *et al.* Colorectal cancer risk reduction following macrogol exposure: a cohort and nested case control study in the UK. *PloS one* 2013;8(12):e83203.
110. Hong JL, Meier CR, Sandler RS, *et al.* Risk of colorectal cancer after initiation of orlistat: matched cohort study. *BMJ* 2013;347:f5039.

111. Makar GA, Holmes JH, Yang YX. Angiotensin-converting enzyme inhibitor therapy and colorectal cancer risk. *Journal of the National Cancer Institute* 2014;106(2):djt374.
112. Smiechowski B, Azoulay L, Yin H, *et al.* The use of metformin and colorectal cancer incidence in patients with type II diabetes mellitus. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2013;22(10):1877-83.
113. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to bisphosphonates and risk of common non-gastrointestinal cancers: series of nested case-control studies using two primary-care databases. *British journal of cancer* 2013;109(3):795-806.
114. Tate AR, Martin AG, Ali A, *et al.* Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer. *BMJ open* 2011;1(1):e000025.
115. Levesque LE, Hanley JA, Kezouh A, *et al.* Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 2010;340:b5087.
116. Tamim H, Monfared AA, LeLorier J. Application of lag-time into exposure definitions to control for protopathic bias. *Pharmacoepidemiology and drug safety* 2007;16(3):250-8.
117. Brookhart MA, Patrick AR, Dormuth C, *et al.* Adherence to lipid-lowering therapy and the use of preventive health services: an investigation of the healthy user effect. *American journal of epidemiology* 2007;166(3):348-54.
118. Shrank WH, Patrick AR, Brookhart MA. Healthy user and related biases in observational studies of preventive interventions: a primer for physicians. *Journal of general internal medicine* 2011;26(5):546-50.
119. Suissa S, Dell'aniello S, Vahey S, *et al.* Time-window bias in case-control studies: statins and lung cancer. *Epidemiology* 2011;22(2):228-31.
120. Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiology and drug safety* 2007;16(3):241-9.
121. Chapman RH, Petrilla AA, Benner JS, *et al.* Predictors of adherence to concomitant antihypertensive and lipid-lowering medications in older adults: a retrospective, cohort study. *Drugs Aging* 2008;25(10):885-92.

122. Setoguchi S. Statins and cancer in the elderly. *Cardiology Review* 2007;24(9):13-16.
123. Boudreau DM, Yu O, Johnson J. Statin use and cancer risk: a comprehensive review. *Expert Opin Drug Saf* 2010;9(4):603-21.
124. Friis S and Olsen J. Statin use and cancer risk: An epidemiologic review. *Cancer Invest* 2006;24(4):413-24.
125. DerSimonian R and Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials* 2007;28(2):105-14.
126. Higgins JP, Thompson SG, Deeks JJ, *et al.* Measuring inconsistency in meta-analyses. *BMJ* 2003;327(7414):557-60.
127. Yang YX, Hennessy S, Prokert K, *et al.* Chronic statin therapy and the risk of colorectal cancer. *Pharmacoepidemiology and Drug Safety* 2008;17(9):869-76.
128. Kaye JA and Jick H. Statin use and cancer risk in the General Practice Research Database. *British Journal of Cancer* 2004;90(3):635-7.
129. Kaye JA, Meier CR, Walker AM, *et al.* Statin use, hyperlipidaemia, and the risk of breast cancer. *British Journal of Cancer* 2002;86(9):1436-9.
130. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to statins and risk of common cancers: A series of nested case-control studies. *BMC cancer* 2011;11(409).
131. Hippisley-Cox J and Coupland C. Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database. *BMJ* 2010;340:c2197.
132. Farwell WR, D'Avolio LW, Scranton RE, *et al.* Statins and prostate cancer diagnosis and grade in a veterans population. *J Natl Cancer Inst* 2011;103(11):885-92.
133. Tan N, Klein EA, Li J, *et al.* Statin use and risk of prostate cancer in a population of men who underwent biopsy. *Journal of Urology* 2011;186(1):86-90.
134. Murtola TJ, Tammela TLJ, Maattanen L, *et al.* Prostate cancer and PSA among statin users in the finnish prostate cancer screening trial. *International Journal of Cancer* 2010;127(7):1650-59.
135. Robertson DJ, Riis AH, Friis S, *et al.* Neither long-term statin use nor atherosclerotic disease is associated with risk of colorectal cancer. *Clinical Gastroenterology & Hepatology* 2010;8(12):1056-61.

136. Woditschka S, Habel LA, Udaltsova N, *et al.* Lipophilic statin use and risk of breast cancer subtypes. *Cancer Epidemiology, Biomarkers & Prevention* 2010;19(10):2479-87.
137. Flick ED, Habel LA, Chan KA, *et al.* Statin use and risk of colorectal cancer in a cohort of middle-aged men in the us: A prospective cohort study. *Drugs* 2009;69(11):1445-57.
138. Hachem C, Morgan R, Johnson M, *et al.* Statins and the risk of colorectal carcinoma: a nested case-control study in veterans with diabetes. *American Journal of Gastroenterology* 2009;104(5):1241-8.
139. Boudreau DM, Yu O, Buist DSM, *et al.* Statin use and prostate cancer risk in a large population-based setting. *Cancer Causes & Control* 2008;19(7):767-74.
140. Boudreau DM, Koehler E, Rulyak SJ, *et al.* Cardiovascular medication use and risk for colorectal cancer. *Cancer Epidemiology, Biomarkers & Prevention* 2008;17(11):3076-80.
141. Farwell WR, Scranton RE, Lawler EV, *et al.* The association between statins and cancer incidence in a veterans population. *Journal of the National Cancer Institute* 2008;100(2):134-9.
142. Friedman GD, Flick ED, Udaltsova N, *et al.* Screening statins for possible carcinogenic risk: up to 9 years of follow-up of 361,859 recipients. *Pharmacoepidemiology and drug safety* 2008;17(1):27-36.
143. Boudreau DM, Yu O, Miglioretti DL, *et al.* Statin use and breast cancer risk in a large population-based setting. *Cancer Epidemiology, Biomarkers & Prevention* 2007;16(3):416-21.
144. Flick ED, Habel LA, Chan KA, *et al.* Statin use and risk of prostate cancer in the California men's health study cohort. *Cancer Epidemiology Biomarkers and Prevention* 2007;16(11):2218-25.
145. Murtola TJ, Tammela TLJ, Lahtela J, *et al.* Cholesterol-lowering drugs and prostate cancer risk: a population-based case-control study. *Cancer Epidemiology, Biomarkers & Prevention* 2007;16(11):2226-32.
146. Setoguchi S, Glynn RJ, Avorn J, *et al.* Statins and the risk of lung, breast, and colorectal cancer in the elderly. *Circulation* 2007;115(1):27-33.
147. Friis S, Poulsen AH, Johnsen SP, *et al.* Cancer risk among statin users: a population-based cohort study. *International Journal of Cancer* 2005;114(4):643-7.

148. Graaf MR, Beiderbeck AB, Egberts ACG, *et al.* The risk of cancer in users of statins. *Journal of Clinical Oncology* 2004;22(12):2388-94.
149. Beck P, Wysowski DK, Downey W, *et al.* Statin use and the risk of breast cancer. *Journal of clinical epidemiology* 2003;56(3):280-5.
150. Blais L, Desgagne A, LeLorier J. 3-Hydroxy-3-methylglutaryl coenzyme A reductase inhibitors and the risk of cancer: a nested case-control study. *Archives of internal medicine* 2000;160(15):2363-8.
151. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;127(8 Pt 2):757-63.
152. Jick H, Jick S, Derby LE, *et al.* Calcium-channel blockers and risk of cancer. *Lancet* 1997;349(9051):525-8.
153. Tan N, Klein EA, Li J, *et al.* Statin use and risk of prostate cancer in a population of men who underwent biopsy. *The Journal of urology* 2011;186(1):86-90.
154. Garcia Rodriguez LA and Gonzalez-Perez A. Inverse association between nonsteroidal anti-inflammatory drugs and prostate cancer. *Cancer Epidemiology Biomarkers and Prevention* 2004;13(4):649-53.
155. Garcia Rodriguez LA and Huerta-Alvarez C. Reduced incidence of colorectal adenoma among long-term users of nonsteroidal antiinflammatory drugs: a pooled analysis of published studies and a new population-based study. *Epidemiology* 2000;11(4):376-81.
156. Hippisley-Cox J, Vinogradova Y, Coupland C, *et al.* Risk of malignancy in patients with schizophrenia or bipolar disorder: Nested case-control study. *Archives of general psychiatry* 2007;64(12):1368-76.
157. Perron L, Bairati I, Harel F, *et al.* Antihypertensive drug use and the risk of prostate cancer (Canada). *Cancer causes & control : CCC* 2004;15(6):535-41.
158. Hudson M, Rahme E, Richard H, *et al.* Comparison of measures of medication persistency using a prescription drug database. *American heart journal* 2007;153(1):59-65.
159. Andrade SE, Kahler KH, Frech F, *et al.* Methods for evaluation of medication adherence and persistence using automated databases. *Pharmacoepidemiol Drug Saf* 2006;15(8):565-74; discussion 75-7.
160. Caetano PA, Lam JM, Morgan SG. Toward a standard definition and measurement of persistence with drug therapy: Examples from research on statin

- and antihypertensive utilization. *Clinical therapeutics* 2006;28(9):1411-24; discussion 10.
161. Hess LM, Raebel MA, Conner DA, *et al.* Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. *The Annals of pharmacotherapy* 2006;40(7-8):1280-88.
162. Simpson SH, Eurich DT, Majumdar SR, *et al.* A meta-analysis of the association between adherence to drug therapy and mortality. *BMJ* 2006;333(7557):15.
163. Ronquist G, Garcia Rodriguez LA, Ruigomez A, *et al.* Association between Captopril, Other Antihypertensive Drugs and Risk of Prostate Cancer. *Prostate* 2004;58(1):50-56.
164. Cyrus-David MS, Weinberg A, Thompson T, *et al.* The effect of statins on serum prostate specific antigen levels in a cohort of airline pilots: a preliminary report. *The Journal of urology* 2005;173(6):1923-5.
165. Cyrus-David MS, Weinberg A, Thompson T, *et al.* The effect of statins on serum prostate specific antigen levels in a cohort of airline pilots: a preliminary report. *Journal of Urology* 2005;173(6):1923-5.
166. Collin SM, Martin RM, Metcalfe C, *et al.* Prostate-cancer mortality in the USA and UK in 1975-2004: an ecological study. *The lancet oncology* 2008;9(5):445-52.
167. Chan TF, Wu CH, Lin CL, *et al.* Statin use and the risk of breast cancer: a population-based case-control study. *Expert opinion on drug safety* 2014;13(3):287-93.
168. Lutski M, Shalev V, Porath A, *et al.* Continuation with statin therapy and the risk of primary cancer: a population-based study. *Preventing chronic disease* 2012;9:E137.
169. Bjorkhem-Bergman L, Backheden M, Soderberg Lofdal K. Statin treatment reduces the risk of hepatocellular carcinoma but not colon cancer-results from a nationwide case-control study in Sweden. *Pharmacoepidemiology and drug safety* 2014;23(10):1101-6.
170. Cheng MH, Chiu HF, Ho SC, *et al.* Statin use and the risk of colorectal cancer: a population-based case-control study. *World journal of gastroenterology : WJG* 2011;17(47):5197-202.
171. Clancy Z, Keith SW, Rabinowitz C, *et al.* Statins and colorectal cancer risk: a longitudinal study. *Cancer causes & control : CCC* 2013;24(4):777-82.

172. Lakha F, Theodoratou E, Farrington SM, *et al.* Statin use and association with colorectal cancer survival and risk: case control study with prescription data linkage. *BMC cancer* 2012;12:487.
173. Cheng MH, Chiu HF, Ho SC, *et al.* Statin use and the risk of female lung cancer: a population-based case-control study. *Lung Cancer* 2012;75(3):275-9.
174. Jespersen CG, Norgaard M, Friis S, *et al.* Statin use and risk of prostate cancer: a Danish population-based case-control study, 1997-2010. *Cancer epidemiology* 2014;38(1):42-7.
175. Lustman A, Nakar S, Cohen AD, *et al.* Statin use and incident prostate cancer risk: does the statin brand matter? A population-based cohort study. *Prostate Cancer Prostatic Dis* 2014;17(1):6-9.
176. Nordstrom T, Clements M, Karlsson R, *et al.* The risk of prostate cancer for men on aspirin, statin or antidiabetic medications. *Eur J Cancer* 2015;51(6):725-33.
177. Williams T, van Staa T, Puri S, *et al.* Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Therapeutic advances in drug safety* 2012;3(2):89-99.
178. National Cancer Data Repository. from [http://www.ncin.org.uk/collecting\\_and\\_using\\_data/national\\_cancer\\_data\\_repository/](http://www.ncin.org.uk/collecting_and_using_data/national_cancer_data_repository/).
179. Chisholm J. The Read clinical classification. *BMJ* 1990;300(6732):1092.
180. Clinical Practice Research Datalink (CPRD). from <http://www.cprd.com>.
181. Cancer Research UK. from <http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared#ref-0>.
182. Cancer Statistics Registrations, England (Series MB1), No. 31, 2000 - No. 41, 2010; <http://www.ons.gov.uk/ons/index.html>.
183. Mortality Statistics: Deaths Registered in England and Wales (Series DR), 2013. from <http://www.ons.gov.uk/ons/rel/vsob1/mortality-statistics--deaths-registered-in-england-and-wales--series-dr-/2013/index.html>.
184. CPRD HSCIC - Patient based linkage. Retrieved 19 May, 2015, from <http://www.cprd.com/researchpractice/researchgppractice.asp>.
185. HSCIC. A Guide to Linked Mortality Data from Hospital Episode Statistics and the Office for National Statistics.

186. Mihaylova B, Emberson J, Blackwell L, *et al.* The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *Lancet* 2012;380(9841):581-90.
187. Charlton R, Snowball J, Bloomfield K, *et al.* Colorectal cancer incidence on the General Practice Research Database. *Pharmacoepidemiology and drug safety* 2012;21(7):775-83.
188. Haynes K, Forde KA, Schinnar R, *et al.* Cancer incidence in The Health Improvement Network. *Pharmacoepidemiology and drug safety* 2009;18(8):730-6.
189. Mitropoulos KA, Armitage JM, Collins R, *et al.* Randomized placebo-controlled study of the effects of simvastatin on haemostatic variables, lipoproteins and free fatty acids. The Oxford Cholesterol Study Group. *European heart journal* 1997;18(2):235-41.
190. Pascoe SW, Neal RD, Heywood PL, *et al.* Identifying patients with a cancer diagnosis using general practice medical records and Cancer Registry data. *Family practice* 2008;25(4):215-20.
191. Tate AR, Nicholson A, Cassell JA. Are GPs under-investigating older patients presenting with symptoms of ovarian cancer? Observational study using General Practice Research Database. *British journal of cancer* 2010;102(6):947-51.
192. O'Dowd EL, McKeever TM, Baldwin DR, *et al.* What characteristics of primary care and patients are associated with early death in patients with lung cancer in the UK? *Thorax* 2015;70(2):161-8.
193. Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International journal of epidemiology* 2015.
194. Boggon R, van Staa TP, Chapman M, *et al.* Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiology and drug safety* 2013;22(2):168-75.
195. Lewis JD, Bilker WB, Weinstein RB, *et al.* The relationship between time since registration and measured incidence rates in the General Practice Research Database. *Pharmacoepidemiology and drug safety* 2005;14(7):443-51.
196. 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.



197. Clogg CC, Petkova E, Haritou A. Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology* 1995;100(5):1261-93.
198. Collins R, Armitage J, Parish S, *et al.* Effects of cholesterol-lowering with simvastatin on stroke and other major vascular events in 20536 people with cerebrovascular disease or other high-risk conditions. *Lancet* 2004;363(9411):757-67.
199. Lewis JD, Bilker WB, Weinstein RB, *et al.* The relationship between time since registration and measured incidence rates in the General Practice Research Database. *Pharmacoepidemiology & Drug Safety* 2005;14(7):443-51.
200. Austin PC. A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. *Multivariate behavioral research* 2011;46(1):119-51.
201. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics* 2010;25(1):1-21.
202. White IR and Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine* 2010;29(28):2920-31.
203. Royston P. Multiple imputation of missing values. *Stata Journal* 2004;4(3):227-41.
204. Sterne JA, White IR, Carlin JB, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
205. Chapman RH, Petrilla AA, Benner JS, *et al.* Predictors of adherence to concomitant antihypertensive and lipid-lowering medications in older adults: a retrospective, cohort study. *Drugs & aging* 2008;25(10):885-92.
206. Douglas I, Evans S, Smeeth L. Effect of statin treatment on short term mortality after pneumonia episode: cohort study. *BMJ* 2011;342:d1642.
207. Boudreau DM, Yu O, Johnson J. Statin use and cancer risk: a comprehensive review. *Expert opinion on drug safety* 2010;9(4):603-21.
208. Schneeweiss S, Patrick AR, Sturmer T, *et al.* Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Medical care* 2007;45(10 Supl 2):S131-42.

209. Klein R, Klein BE, Moss SE, *et al.* Association of ocular disease and mortality in a diabetic population. *Arch Ophthalmol* 1999;117(11):1487-95.
210. Lee DJ, Gomez-Marin O, Lam BL, *et al.* Glaucoma and survival: the National Health Interview Survey 1986-1994. *Ophthalmology* 2003;110(8):1476-83.
211. Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. *Therapeutic advances in drug safety* 2013;4(2):53-62.
212. Lash TL, Fox MP, MacLehose RF, *et al.* Good practices for quantitative bias analysis. *International journal of epidemiology* 2014;43(6):1969-85.
213. Liu CJ and Hu YW. Immortal time bias in retrospective analysis: is there a survival benefit in patients with glioblastoma who received prolonged treatment of adjuvant valganciclovir? *International journal of cancer. Journal international du cancer* 2014;135(1):250-1.
214. Clancy Z, Keith SW, Rabinowitz C, *et al.* Statins and colorectal cancer risk: a longitudinal study. *Cancer Causes & Control* 2013;24(4):777-82.
215. Madigan D, Ryan PB, Schuemie M, *et al.* Evaluating the impact of database heterogeneity on observational study results. *American journal of epidemiology* 2013;178(4):645-51.

## 8 Appendix A - Supplementary materials for Chapter 2

### Questionnaire A.1: Code list questionnaire

#### PART 1 – SUMMARY INFORMATION

- 1** In deciding which OXMIS/Read codes to include in your case definition(s), which of the following strategies did you employ? (please indicate with a **X**, all that apply)

Key word/synonym search

Utilisation of a code list from a previous study

Consultation with a GP or health professional

Other strategy/resource (please describe)

- 2** How many researchers substantively participated in the development/review of the code list, and what is/are their specific professional background(s)?

- 3** How were codes for borderline or suspected malignancies dealt with? (Please indicate, with a **X**, which apply)

Borderline codes were included along with “definite” codes

Borderline codes were excluded from the code list

Borderline codes identified and individually reviewed/ validated later

Borderline codes separately included in a sensitivity/secondary analysis

Other (please describe)

#### PART 2 – SPECIFIC DETAILS

To help us usefully summarize how researchers to date have gone about developing code lists for cancer studies, and how the final lists themselves vary, we would be grateful if you would be willing to share some specific details\* on:

**(i) the strategy you used to develop your code list: please briefly describe the overall process, significant updates from previous study code lists and give the keywords for any search. (Include details as a separate attachment if preferred).**

**(ii) your final OXMIS/Read code list for identification of cases. Please insert list below or attach separately with your reply\*.**

**(iii) your borderline codes list. Please insert list below or attach separately with your reply**

**Table 8.1: Summary description table of included studies**

	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods*	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
1	Hippisley-Cox 2013[1]	QRESEARCH	2000-2012	Symptoms	Colorectal Prostate	None stated	None stated	None	None stated	None stated	≥1 year registration with GP	Diagnosis of incident cancer within 2 years after study entry using Read codes and/or ICD-9 (linked ONS data).
2	Hippisley-Cox 2013[2]	QRESEARCH	2000-2012	Symptoms	Breast Colorectal	None stated	None stated	None	None stated	None stated	≥1 year registration with GP	Diagnosis of incident cancer within 2 years after study entry using Read codes and/or ICD-9 (linked ONS data).
3	Osborn 2013[3]	THIN	1990-2008	None-incidence	Breast Colorectal Prostate	None stated	None stated	None	None stated	Previous diagnosis of cancer of interest	≥6 months of follow up	Diagnosis of incident cancer (invasive and in situ) among patients with ≥6 months of follow up.
4	Vinogradova 2013[4]	QRESEARCH + CPRD	1997-2011	Oral Bisphosphonates	Colorectal	None stated	None stated	None	None stated	None stated	≥2 years follow-up	Diagnosis of incident cancer among patients with ≥2 years follow-up.
5	Azoulay 2012[5]	CPRD	1995-2008	Angiotensin Receptor Blockers	Breast Colon Prostate	None stated	None stated	None	None stated	Previous diagnosis of cancer of interest	≥2 years follow-up	Diagnosis of incident cancer among patients with ≥2 years follow-up
6	Bhaskaran 2012[6]	CPRD	1995-2010	Angiotensin Receptor Blockers	Breast Colon Prostate	Stated in publication*	None stated	None	None stated	Previous diagnosis of any cancer. Exclusion of borderline and suspected malignancy codes.	≥1 year follow-up	Diagnosis of incident cancer among patients with ≥1 year follow-up. Codes for borderline or suspected malignancies were excluded from the final code list.
7	Boggon 2012[7]	CPRD	1997-2006	Diabetes and related therapy	Breast Colorectal Prostate	None stated	None stated	Linkage to UK cancer registry, Hospital Episode Statistics, and free-text	None stated	None stated	None stated	Diagnosis of incident cancer. In-situ or cases with non-malignant ICD-10 codes were defined as 'other type'.
8	Cardwell 2012[8]	CPRD	1996-2006	Oral Bisphosphonates	Breast Colorectal Prostate	Stated in publication*	None stated	None	None stated	Previous diagnosis of any cancer	≥3 years UTS medical records	Diagnosis of incident cancer after ≥3-years initial follow-up.
9	Collins 2012[9]	THIN	2000-2008	Symptoms	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of colorectal cancer	≥1 year UTS medical records	Diagnosis of colorectal cancer after ≥1-year initial follow-up.
10	Dregan 2012[10]	CPRD	2001-2007	Symptoms	Colorectal	Stated in publication*	Web appendix	Linkage to UK cancer registry	None stated	None stated	≥1 year follow-up	Diagnosis of cancer after ≥1 year UTS follow-up. Patients were excluded if they had a previous diagnosis of cancer or had records of previous alarm symptoms prior to cohort entry. Inclusion of ICD-10 and ICD-0-2 codes.

[Table 8.1 continued over]

[Table 8.1 continued]												
	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
11	Jick 2012[11]	CPRD	1991-2009	Testosterone therapy	Prostate	None stated	None stated	None	None stated	Previous diagnosis of prostate cancer	≥2 years UTS medical records	Diagnosis of prostate cancer after ≥2 years initial follow-up. Patients with an indication of pre-existing prostate cancer prior to cohort entry were excluded.
12	Mackenzie 2012[12]	CPRD	1987-2010	Spironolactone	Breast	Stated in publication*	Provided within publication	None	None stated	Previous diagnosis of breast cancer	None stated	Diagnosis of breast cancer in patients with UTS follow-up. Both invasive and in-situ cases were included.
13	Qiu 2012[13]	CPRD	1995-2008	Metformin & sulphonylurea	Breast Colorectal Prostate	None stated	None stated	None	None stated	None stated	≥1 year of electronic medical records	Diagnosis of invasive solid tumours in patients with ≥1 year UTS follow-up after index date.
14	Redaniel 2012[14]	CPRD	1987-2007	Diabetes and related treatment	Breast	None stated	None stated	None	None stated	Previous diagnosis of breast cancer	≥1 year electronic medical records	Diagnosis of breast cancer among patients with ≥1 year UTS follow-up after index date.
15	Singh 2012[15]	CPRD	1987-1992	Epilepsy	Breast Colorectal Prostate	None stated	None stated	None	None stated	Previous diagnosis of any cancer	None stated	Diagnosis of cancer after the cohort entry or before diagnosis of epilepsy.
16	Bodmer 2012[16]	CPRD	1995-2009	Metformin	Colorectal	None stated	None stated	None	None stated	History of any cancer	≥3 years of follow-up	Diagnosis of colorectal cancer in patients with ≥3 years active history prior to diagnosis date.
17	Charlton 2012[17]	CPRD	2007	None incidence study	Colorectal	None stated	None stated	1. Diagnostic algorithm 2. Free-text 3. Comparison of rates	Supporting evidence of diagnosis <sup>1</sup>	Previous diagnosis of colorectal cancer	≥6 months UTS records	Diagnosis of colorectal cancer in patients with ≥6 months UTS follow-up. Codes including malignant neoplasms, neoplasms of uncertain behaviour, general codes were used to identify cases.
18	Hippisley-Cox 2012[18]	QRESEARCH	2000-2010	Symptoms	Colorectal	None stated	None stated	None	None stated	History of colorectal cancer	≥2 years of follow up	Diagnosis of cancer within 2 years after study entry using Read codes or ICD-9 (linked ONS data).
20	Vinogradova 2012[20]	QRESEARCH	1996-2011	Bisphosphonates	Breast Colorectal Prostate	None stated	Read dictionary: B chapter	None	None stated	1.Previous diagnosis of any cancer; 2.Secondary cancers (B56-58); 3.Previous mastectomy	≥2 years of electronic medical records	Diagnosis of incident cancer in patients with ≥2 years medical records prior to date of diagnosis. Female patients with a prior record of a mastectomy were excluded.
21	Walker AJ 2011[21]	CPRD	Not stated	Tricyclic antidepressants	Breast Colorectal Prostate	None stated	Available on request	None	None stated	Previous diagnosis of any cancer	≥5 years of follow-up	Diagnosis of colorectal cancer in patients with ≥5 years follow-up.

[Table 8.1 continued over]

[Table 8.1 continued]												
	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
19	Van Staa 2012[19]	CPRD	1997-2006	Glucose lowering drugs	Breast Colorectal Prostate	None stated	None stated	None	None stated	None stated	None stated	Diagnosis of incident cancer.
22	Green 2011[22]	CPRD	1995-2005	Hormone therapy	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of gastrointestinal cancer	≥1 year of follow-up	Diagnosis of gastrointestinal cancer in female patients with ≥1 year follow-up.
23	Azoulay 2011[23]	CPRD	1988-2009	Metformin	Prostate	None stated	None stated	Diagnostic algorithm	Supporting evidence of diagnosis <sup>2</sup>	Previous diagnosis of prostate cancer	≥1 year UTS medical records	Diagnosis of prostate cancer among male patients with ≥1 year follow-up identified by an algorithm. The algorithm included medical codes for prostate cancer as well as codes for prostate biopsies, surgeries, radiation therapy, and androgen deprivation therapy.
24	Azoulay 2011[24]	CPRD	1988-2007	Antipsychotics	Breast	None stated	None stated	Diagnostic algorithm	Supporting evidence of diagnosis <sup>3</sup>	Previous diagnosis of breast cancer	≥1 year UTS medical records	Diagnosis of breast cancer among female patients with ≥1 year follow-up identified by an algorithm. The algorithm included medical codes for breast cancer as well as codes for mastectomies, lumpectomies, axillary node dissection, oncologist consultation, chemo-radiotherapy, and post-operative hormone therapy.
25	Damery 2011[25]	THIN	2000-2006	Iron deficiency anaemia	Colorectal	None stated	None stated	None	None stated	None stated	None stated	Diagnosis of colorectal cancer in patients with ≥1 year follow-up.
26	Suissa 2011[26]	CPRD	2002-2006	Insulin Glargine	Breast	None stated	None stated	Diagnostic algorithm	Supporting evidence of diagnosis <sup>3</sup>	Previous diagnosis of breast cancer	≥1 year UTS medical records	Diagnosis of breast cancer among female patients with ≥1 year UTS follow-up identified by an algorithm (see #24 for details).
27	Marshall 2011[27]	THIN	2001-2006	Symptoms	Colorectal	None stated	Available on request	None	None stated	None stated	≥2 years of electronic medical records	Diagnosis of colorectal cancer in patients with ≥2 years electronic records prior to date of diagnosis.
28	Vinogradova 2011[28]	QRESEARCH	1998-2008	Statins	Breast Colorectal Prostate	None stated	Read dictionary: B chapter	None	None stated	1. Previous diagnosis of any cancer; 2. Secondary cancers (B56-58); 3. Previous mastectomy, or tamoxifen prescription	≥6 years of electronic medical records	Diagnosis of incident cancer in patients with ≥6 years of medical records prior to date of diagnosis. Female patients with a prior record of a mastectomy or prescription for tamoxifen were excluded.

[Table 8.1 continued over]

[Table 8.1 continued]

	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
29	Vinogradova 2011[29]	QRESEARCH	1997-2008	Cyclooxygenase-2 inhibitors	Breast Colorectal Prostate	None stated	Read dictionary: B chapter	None	None stated	1. Previous diagnosis of any cancer; 2. Secondary cancers (B56-58); 3. Previous mastectomy, or tamoxifen prescription	≥6 years electronic medical records	Diagnosis of incident cancer in patients with ≥6 years of medical records prior to date of diagnosis. Female patients with a prior record of a mastectomy or tamoxifen use ≥12 months were excluded.
30	Hippisley-Cox 2010[30]	QRESEARCH	2002-2008	Statins	Breast Colon Prostate	None stated	Available on request	None	None stated	Previous diagnosis of outcome of interest	≥1 year registration with GP	Diagnosis of incident cancer in patients with ≥1 year registration with GP.
31	Becker 2010[31]	CPRD	1994-2005	Parkinson's Disease	Breast Colorectal Prostate	None stated	None stated	Manual Review (random sample of one third of cases)	Chemotherapy, radiotherapy or cancer related surgery	Previous diagnosis of any cancer	≥3 years electronic medical records	Diagnosis of incident cancer in patients with ≥3 years of medical records prior to cohort entry. Validation of one-third of the cases required patients to have related surgery, or chemo-radiotherapy around the date of diagnosis
32	Schneider 2010[32]	CPRD	1995-2005	Chronic obstructive pulmonary disease	Breast	None stated	None stated	None	None stated	Previous history of any cancer	≥3 years electronic medical records	Diagnosis of incident cancer in patients with ≥3 years medical records.
33	Green 2010[33]	CPRD	1995-2005	Oral Bisphosphonates	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of gastrointestinal cancer	≥1 year of follow-up	Diagnosis of incident gastrointestinal cancer in patients with ≥1 year follow-up.
34	Bodmer 2010[34]	CPRD	1994-2005	Diabetes-Metformin	Breast	None stated	None stated	Manual Review	Supportive evidence of diagnosis <sup>3</sup>	History of any cancer	≥3 years of electronic medical records	Diagnosis of incident breast cancer (invasive or in situ) followed by breast surgery, chemo-radiotherapy, or antiestrogen therapy in patients with ≥3 years recorded medical records.
35	Armstrong 2010[35]	CPRD	1987-2001	Inflammatory Bowel Disease	Breast	None stated	None stated	None	None	History of any cancer	≥1 year of electronic medical records	Diagnosis of incident cancer (malignant diagnosis code) in patients with ≥1 year of medical records. Patients with non-specific codes only were not considered as cases.
36	Currie 2009[36]	THIN	2000 onwards	Glucose lowering drugs	Breast Colorectal Prostate	None stated	None stated	None	None stated	Previous diagnosis of any cancer	≥6 months medical history	First record of any solid tumour in patients with ≥6 months medical history.

[Table 8.1 continued over]



[Table 8.1 continued]												
	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
37	Schneider 2009[37]	CPRD	Up to 2007	HRT	Breast	None stated	None stated	Diagnostic algorithm	Chemotherapy, radiotherapy or cancer related surgery	Previous diagnosis of any cancer	None stated	Diagnosis of incident gynaecological cancer in female patients.
38	Haynes 2009[38]	THIN	1992-2007	None – Incidence study	Colorectal	None stated	Available on request	Comparison of rates	None	Previous diagnosis of any cancer	Patients with unacceptable registration status excluded <sup>4</sup>	Diagnosis of incident cancer in patients with an acceptable status of registration. Read codes were consistent with ICD-10 codes used by the UK cancer registry. Patients with codes for history of cancer, or cancer related treatment, and neoplasms of uncertain significance were added in a sensitivity analysis
39	Simon 2009[39]	CPRD	1987-2001	Rheumatoid arthritis	Breast Colorectal	Stated in publication*	None stated	None	None stated	None stated	None stated	Diagnosis of incident cancer.
40	van Staa 2009[40]	CPRD + THIN	Not stated	Testosterone therapy	Breast	None stated	None stated	None	None stated	None stated	None stated	Diagnosis of incident breast cancer (ICD9, C74)
41	Brauchli 2009[41]	CPRD	1994-2005	Psoriasis	Breast Colorectal Prostate	None stated	None stated	1. Manual Review 2. Diagnostic algorithm	Radiotherapy; chemotherapy; endocrine therapy; referral to specialist, surgery; hospitalised; died within 180 days of diagnosis	History of any cancer	≥3 years of electronic medical records	Diagnosis of incident cancer (malignant or in situ). Patients with evidence of diagnosis such as surgery, chemo-radiotherapy, endocrine therapy, referred to a specialist, or died within 180 days after diagnosis.
42	Smeeth 2008[42]	THIN	1995-2006	Statins	Breast Prostate	None stated	None stated	None	None stated	History of any cancer	≥1 year continuous registration with GP	Diagnosis of cancer in patients with ≥1 year continuous registration with their GP.
43	Opatrny 2008[43]	CPRD	1988-2004	HRT	Breast	None stated	None stated	Diagnostic algorithm	Supporting evidence of diagnosis <sup>3</sup>	Previous diagnosis of breast cancer	Member of UTS practice	Case definition same as #24
44	Weiner 2008[44]	CPRD	1990-1999	HRT	Breast Colorectal	None stated	None stated	None	None stated	1. History of any cancer; 2. Previous hysterectomy	None stated	Diagnosis of incident cancer among female patients.
45	Lewis 2008[45]	CPRD	1987-2002	Dermatitis Herpetiformis	Breast	None stated	None stated	None	None stated	None stated	≥1 year CPRD records	Diagnosis of incident cancer among female patients with ≥1 year CPRD records.
46	van Staa 2008[46]	CPRD	Not stated	Hormone therapy	Colorectal Breast	None stated	None stated	None	None stated	None stated	None stated	Diagnosis of incident colorectal cancer among female patients.

[Table 8.1 continued over]

[Table 8.1 continued]												
	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
47	Yang 2008[47]	CPRD	1987-2002	Statins	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of colorectal cancer	≥5 years of follow up	Diagnosis of incident colorectal cancer among patients with ≥5 years continuous follow-up in the CPRD.
48	Lewis 2007[48]	THIN	1986-2003	Aspirin	Colon	None stated	Available on request	None	None stated	Previous diagnosis of colon cancer	≥1 year GP registration	Diagnosis of incident colon cancer among patients with ≥1 year registration with their GP.
49	Parker 2007[49]	QRESEARCH	1998-2003	Rectal and postmenopausal bleeding	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of cancer of interest	≥1 year GP registration	Diagnosis of incident cancer among patients with ≥1 year registration with their GP.
50	Jones 2007[50]	CPRD	1994-2000	Alarm symptoms	Colorectal	None stated	Available on request	Manual review	Unspecified	Previous diagnosis of any cancer	None stated	Diagnosis of incident cancer.
51	Srinivasan 2007[51]	CPRD	Not stated	HRT	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of any cancer	≥1 year UTS follow-up	Diagnosis of incident cancer after ≥1 year UTS follow-up.
52	Jackson 2007[52]	CPRD	1987-2002	Ursodeoxycholic acid	Breast	None stated	None stated	None	None stated	None stated	≥1 year UTS follow-up	Diagnosis of incident cancer after ≥1 year UTS follow-up.
53	Le Jeune 2007[53]	THIN	Up to 2004	Idiopathic pulmonary fibrosis	Breast Prostate	None stated	None stated	None	None stated	None stated	≥1 year follow-up	Diagnosis of incident cancer after ≥1 year follow-up.
54	Hippisley-Cox 2007[54]	QRESEARCH	1995-2005	Schizophrenia	Breast Colon Prostate	None stated	None stated	None	None stated	Previous mastectomy, or tamoxifen prescription prior to first record of cancer	≥1 year computerised records	Diagnosis of incident cancer after ≥1 year follow-up. Patients with previous mastectomies or tamoxifen prescription were excluded.
55	Tannen 2007[55]	CPRD	1990-1999	HRT	Breast Colorectal	None stated	None stated	None	None stated	History of any cancer; previous hysterectomy	None stated	Diagnosis of incident cancer among female patients.
56	Tannen 2007[56]	CPRD	1990-1998	Estrogen	Breast Colorectal	None stated	None stated	None	None stated	History of any cancer; previous hysterectomy	None stated	Same case definition as #55.
57	Vinogradova 2007[57]	QRESEARCH	1995-2005	Statins, NSAIDs & Cyclooxygenase-2 Inhibitors	Colorectal	None stated	None stated	Comparison of rates	None stated	Previous diagnosis of any cancer	≥6 years of electronic medical records	Diagnosis of incident colorectal cancer.
58	Yang 2007[58]	CPRD	1987-2002	Proton Pump Inhibitors	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of colorectal cancer	≥5 years UTS follow-up	Diagnosis of incident colorectal cancer among patients with ≥5 year continuous follow-up in the CPRD.
59	Gonzalez-Perez 2006[59]	CPRD	1994-2001	Asthma	Breast Colorectal Prostate	None stated	None stated	1. Manual Review 2. Questionnaire to GP	Unspecified	Previous diagnosis of any cancer	≥1 year enrolment with GP and ≥1 since first prescription	Diagnosis of incident cancer among patients with ≥1 year enrolment with GP.
[Table 8.1 continued over]												

[Table 8.1 continued]												
	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
60	Gonzalez-Perez 2005[60]	CPRD	1995-2001	Antidepressants	Breast	None stated	None stated	1. Manual Review 2. Q to GP	Supportive evidence of diagnosis <sup>3</sup>	Previous diagnosis of any cancer	≥1 year enrolment with GP and ≥1 year since first prescription	Diagnosis of incident breast cancer among patients with ≥1 year enrolment with GP. Requirement of supportive evidence of diagnosis.
61	Gonzalez-Perez 2005[61]	CPRD	1995-2001	Diabetes Mellitus	Prostate	None stated	None stated	1. Manual Review 2. Questionnaire to GP	Unspecified	Previous diagnosis of any cancer	≥1 year enrolment with GP and ≥1 year since first prescription	Diagnosis of incident prostate cancer among patients with ≥1 year enrolment with GP.
62	Bradbury 2005[62]	CPRD	1991-2001	Obesity	Prostate	None stated	None stated	None	Treatment or additional diagnosis of prostate cancer within 6 months of diagnosis	Previous diagnosis of any cancer	≥1 year computerised records	Diagnosis of incident prostate cancer among patients with ≥1 year computerised records. Treatment or follow-up visits within 6 months of diagnosis.
63	Kaye 2005[63]	CPRD	1987-2002	Antibiotics	Breast	None stated	None stated	None	Treatment unspecified	Previous diagnosis of any cancer	≥6 years electronic medical records	Diagnosis of incident breast cancer among patients with ≥6 months recorded history.
64	Lewis 2005[64]	CPRD	1987-2003	None – incidence study	Breast Colon Prostate	None stated	None stated	None	None stated	Previous diagnosis of any cancer	Up to 3 years minimum registration with GP	Diagnosis of incident cancer among patients with varying durations of history after registration with GP.
65	Garcia Rodriguez 2005[65]	CPRD	1995-2001	Antibiotics	Breast	None stated	None stated	1. Manual Review 2. Questionnaire to GP	Supportive evidence of diagnosis <sup>3</sup>	Previous diagnosis of any cancer	≥1 year enrolment with GP and ≥1 year since first computerised prescription	Diagnosis of incident breast cancer among patients with follow-up ≥1 year enrolment with GP. Diagnoses of invasive and in situ tumours were not distinguished. Additional confirmatory evidence of diagnosis was required to confirm case status.
66	Shao 2005[66]	CPRD	1987-2002	Cholecystectomy	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of colorectal cancer	≥ 1 year UTS follow-up	Diagnosis of incident colorectal cancer among patients with follow-up ≥1 year UTS follow-up.
67	van Staa 2005[67]	CPRD	1987-2001	5-Aminosalicylates	Colorectal	None stated	None stated	Questionnaire to GP	None stated	History of colorectal cancer	None stated	Diagnosis of incident colorectal cancer according to ICD9 (154, 159) classification.
68	Yang 2005[68]	CPRD	1987-2002	Type 2 diabetes	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of colorectal cancer	≥ 1 year UTS follow-up	Diagnosis of incident colorectal cancer following ≥ 1 year UTS follow-up.
69	Gonzalez-Perez 2004[69]	CPRD	1995-2001	Antihypertensive medications	Breast	None stated	None stated	Manual Review	Supportive evidence of diagnosis <sup>3</sup>	Previous diagnosis of any cancer	≥1 year enrolment with GP and ≥1 year since first prescription	Same case definition as #60
70	Ronquist 2004[70]	CPRD	1995-1999	Antihypertensive medications	Prostate	None stated	None stated	1. Manual Review 2. Request of records 3. Comparison of rates	None stated	Previous diagnosis of any cancer	≥2 years enrolment with GP and ≥1 year since first prescription	Diagnosis of incident prostate cancer among patients with ≥2 years enrolment with GP.
[Table 8.1 continued over]												

[Table 8.1 continued]												
	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
71	Kaye 2004[71]	CPRD	1990-2002	Statins	Breast Colorectal Prostate	None stated	None stated	None	None stated	Previous diagnosis of any cancer	≥3 years of follow-up	Diagnosis of incident cancer among patients with ≥3 years follow-up.
72	West 2004[72]	CPRD	1987-2002	Coeliac disease	Breast Prostate	None stated	None stated	None	None stated	None stated	None stated	Diagnosis of incident cancer. Read codes were mapped to ICD codes.
73	Garcia Rodriguez 2004[73]	CPRD	1995-2001	Aspirin and other anti-inflammatory drugs	Breast	None stated	None stated	1. Manual Review 2. Questionnaire to GP 3. Comparison of rates	Unspecified	Previous diagnosis of any cancer	≥1 year enrolment with GP and ≥1 year since first prescription	Same case definition as #60
74	Garcia Rodriguez 2004[74]	CPRD	1995-2001	NSAIDs	Prostate	None stated	None stated	1. Manual Review 2. Questionnaire to GP	Unspecified	Previous diagnosis of any cancer	≥1 year enrolment with GP and ≥1 year since first prescription	Same case definition as #61
75	Solaymani-Dodaran 2004[75]	CPRD	Not stated	Barrett's Oesophagus	Colorectal	None stated	None stated	None	None stated	Previous diagnosis of any cancer	Outcomes observed after 1 year from cohort entry	Diagnosis of incident colorectal cancer following ≥ 1 year after cohort entry.
76	Yang 2004[76]	CPRD	1987-2002	Insulin therapy	Colorectal	None stated	None stated	Manual Review	Clinical events (not stated) supportive of colorectal cancer diagnosis	Previous diagnosis of colorectal cancer	≥3 year follow-up	Diagnosis of incident colorectal cancer following ≥ 3 years UTS follow-up. Additional evidence was required to confirm case status.
77	Meier 2002[77]	CPRD	1992-1997	NSAIDs	Breast Colon	None stated	None stated	Manual Review	Required supportive evidence of diagnosis <sup>3</sup> or GP recorded comments indicating "malignant" or "positive histology"	Previous diagnosis of any cancer	≥3 years of electronic medical records	Diagnosis of incident among patients with ≥3 years medical records. Additional evidence of diagnosis was required to confirm case status.
78	Kaye 2002[78]	CPRD	1992-1998	Statins	Breast	None stated	None stated	Manual Review	Unspecified	1. Previous diagnosis of any cancer 2. Diagnoses with breast cancer at death 3. Uncertain diagnosis of breast cancer	None stated	Diagnosis of incident breast cancer. Patients with an uncertain diagnosis, or diagnosis at death were excluded.
[Table 8.1 continued over]												

[Table 8.1 continued]

	Author & Year	Database	Time-period	Exposure	Cancer(s) investigated	Code list: methods	Code list availability stated in publication	Confirmation of cancer(s)	Additional inclusion criteria: cancer related codes	Exclusion criteria: cancer related codes	Follow-up requirements	Case Description
79	Garcia Rodriguez 2001[79]	CPRD	1994-1997	NSAIDs	Colorectal	None stated	None stated	1. Manual review 2. Request of records 3. Comparison of rates	Manual review unspecified External validation: required biopsy specimen to confirm case	Previous diagnosis of colorectal adenoma, familial polyposis or any cancer prior to cohort entry	≥2 years enrolment with GP	Diagnosis of incident colorectal cancer following ≥ 2 years enrolment with GP.
80	Meier 2000[80]	CPRD	1992-1997	ACE inhibitors	Breast	None stated	None stated	1. Manual review 2. Request of records	Supportive evidence <sup>3</sup> ; and or histological analysis	Previous diagnosis of any cancer	≥3 years drug prescription history	Diagnosis of incident breast cancer following among patients with ≥3 years drug prescription history. Cases were grouped into 3 classifications based on additional supportive evidence of diagnosis.: definite, probable and uncertain
81	Kaye 2000[81]	CPRD	1990-1996	None – incidence study	Breast	None stated	None stated	1. Manual Review 2. Comparison of rates	Breast lump or mammographic abnormality prior to breast cancer diagnosis, and post diagnosis related surgery or therapy	Previous diagnosis of breast cancer	None stated	Potential cases were identified from 3 initial cohorts: (i) diagnosis of breast cancer; women prescribed tamoxifen (no diagnosis code); women with a record for breast cancer related surgery with a diagnosis of cancer unspecified. Potential cases were subsequently confirmed as cases by manual review
82	Langman 2000[82]	CPRD	1993-1995	Anti-inflammatory drugs	Breast Colon Rectum Prostate	None stated	None stated	None	None stated	None	None stated	Diagnosis of incident cancer.
83	Garcia Rodriguez 2000[83]	CPRD	1994-1997	NSAIDs	Colorectal	None stated	None stated	1. Manual Review 2. Q to GP	Unspecified	Previous diagnosis of any cancer	≥2 years enrolment with GP	Diagnosis of incident colorectal adenoma among patients with ≥2 years enrolment with GP.
84	Jick 1997[84]	CPRD	1987- unspecified	Calcium Channel Blockers	Breast Prostate Bowel	None stated	None stated	1. Q to GP 2. Request of records	None stated	History of any cancer	≥4 years electronic medical records	Diagnosis of incident cancer among patients with ≥4 years medical records.

<sup>1</sup> Colorectal cancer surgery, chemotherapy, radiotherapy, dukes staging, palliative care, terminal illness; <sup>2</sup> Prostate cancer surgery, prostate biopsy, chemotherapy, radiotherapy, and androgen deprivation therapy

<sup>3</sup> Breast cancer surgery: mastectomies, lumpectomies, axillary node dissections; chemotherapy; radiotherapy; consultations with oncologist; and post-operative hormone therapy

<sup>4</sup> Unacceptable registration status: not permanently registered; out of sequence year of birth or registration date; missing or invalid transfer out date; year of birth missing or invalid; missing sex information

\*At least one of the following was described in the publication: search of Read dictionary; utilisation of a previous code list; code list reviewed by medical professional

## References: Systematic Review - Included Studies

1. Hippisley-Cox, J. and C. Coupland, *Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm*. Br J Gen Pract, 2013. **63**(606): p. 1-10.
2. Hippisley-Cox, J. and C. Coupland, *Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm*. Br J Gen Pract, 2013. **63**(606): p. 11-21.
3. Osborn, D.P., et al., *Relative incidence of common cancers in people with severe mental illness. Cohort study in the United Kingdom THIN primary care database*. Schizophr Res, 2013. **143**(1): p. 44-9.
4. Vinogradova, Y., C. Coupland, and J. Hippisley-Cox, *Exposure to bisphosphonates and risk of gastrointestinal cancers: series of nested case-control studies with QResearch and CPRD data*. BMJ, 2013. **346**: p. f114.
5. Azoulay, L., et al., *Long-term use of angiotensin receptor blockers and the risk of cancer*. PLoS ONE [Electronic Resource], 2012. **7**(12): p. e50893.
6. Bhaskaran, K., et al., *Angiotensin receptor blockers and risk of cancer: cohort study among people receiving antihypertensive drugs in UK General Practice Research Database*. BMJ, 2012. **344**: p. e2697.
7. Boggon, R., et al., *Cancer recording and mortality in the General Practice Research Database and linked cancer registries*. Pharmacoepidemiol Drug Saf, 2012.
8. Cardwell, C.R., et al., *Exposure to oral bisphosphonates and risk of cancer*. International Journal of Cancer, 2012. **131**(5): p. E717-25.
9. Collins, G.S. and D.G. Altman, *Identifying patients with undetected colorectal cancer: an independent validation of Qcancer (Colorectal)*. Br J Cancer, 2012. **107**(2): p. 260-5.
10. Dregan, A., et al., *Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study*. Cancer Epidemiol, 2012. **36**(5): p. 425-9.
11. Jick, S.S. and K.W. Hagberg, *The risk of adverse outcomes in association with use of testosterone products: a cohort study using the UK-based general practice research database*. Br J Clin Pharmacol, 2013. **75**(1): p. 260-70.
12. Mackenzie, I.S., et al., *Spirolactone and risk of incident breast cancer in women older than 55 years: retrospective, matched cohort study*. BMJ, 2012. **345**: p. e4447.
13. Qiu, H., et al., *Initial metformin or sulphonylurea exposure and cancer occurrence among patients with type 2 diabetes mellitus*. Diabetes Obes Metab, 2013. **15**(4): p. 349-57.

14. Redaniel, M.T., et al., *Associations of type 2 diabetes and diabetes treatment with breast cancer risk and mortality: a population-based cohort study among British women*. *Cancer Causes Control*, 2012. **23**(11): p. 1785-95.
15. Singh, G., et al., *Cancer risk in people with epilepsy using valproate-sodium*. *Acta Neurologica Scandinavica*, 2012. **125**(4): p. 234-40.
16. Bodmer, M., et al., *Use of metformin is not associated with a decreased risk of colorectal cancer: a case-control analysis*. *Cancer Epidemiology, Biomarkers & Prevention*, 2012. **21**(2): p. 280-6.
17. Charlton, R., et al., *Colorectal cancer incidence on the General Practice Research Database*. *Pharmacoepidemiol Drug Saf*, 2012.
18. Hippisley-Cox, J. and C. Coupland, *Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm*. *British Journal of General Practice*, 2012. **62**(594): p. e29-37.
19. van Staa, T.P., et al., *Glucose-lowering agents and the patterns of risk for cancer: a study with the General Practice Research Database and secondary care data*. *Diabetologia*, 2012. **55**(3): p. 654-65.
20. Vinogradova, Y., C. Coupland, and J. Hippisley-Cox, *Exposure to bisphosphonates and risk of cancer: a protocol for nested case-control studies using the QResearch primary care database*. *BMJ Open*, 2012. **2**(1): p. e000548.
21. Walker, A.J., et al., *Tricyclic antidepressants and the incidence of certain cancers: a study using the GPRD*. *British Journal of Cancer*, 2011. **104**(1): p. 193-7.
22. Green, J., et al., *Menopausal hormone therapy and risk of gastrointestinal cancer: nested case-control study within a prospective cohort, and meta-analysis*. *International Journal of Cancer*, 2012. **130**(10): p. 2387-96.
23. Azoulay, L., et al., *Metformin and the incidence of prostate cancer in patients with type 2 diabetes*. *Cancer Epidemiology, Biomarkers & Prevention*, 2011. **20**(2): p. 337-44.
24. Azoulay, L., et al., *The use of atypical antipsychotics and the risk of breast cancer*. *Breast Cancer Research & Treatment*, 2011. **129**(2): p. 541-8.
25. Damery, S., et al., *Iron deficiency anaemia and delayed diagnosis of colorectal cancer: A retrospective cohort study*. *Colorectal Disease*, 2011. **13**(4): p. e53-e60.
26. Suissa, S., et al., *Long-term effects of insulin glargine on the risk of breast cancer*. *Diabetologia*, 2011. **54**(9): p. 2254-62.
27. Marshall, T., et al., *The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance*. *Gut*, 2011. **60**(9): p. 1242-1248.

28. Vinogradova, Y., C. Coupland, and J. Hippisley-Cox, *Exposure to statins and risk of common cancers: a series of nested case-control studies*. BMC Cancer, 2011. **11**: p. 409.
29. Vinogradova, Y., C. Coupland, and J. Hippisley-Cox, *Exposure to cyclooxygenase-2 inhibitors and risk of cancer: Nested case-control studies*. British Journal of Cancer, 2011. **105**(3): p. 452-459.
30. Hippisley-Cox, J. and C. Coupland, *Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database*. BMJ, 2010. **340**: p. c2197.
31. Becker, C., et al., *Cancer risk in association with Parkinson disease: a population-based study*. Parkinsonism & Related Disorders, 2010. **16**(3): p. 186-90.
32. Schneider, C., et al., *Cancer risk in patients with chronic obstructive pulmonary disease*. Pragmatic and Observational Research, 2010. **1**(1): p. 15-23.
33. Green, J., et al., *Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort*. BMJ, 2010. **341**: p. c4444.
34. Bodmer, M., et al., *Long-term metformin use is associated with decreased risk of breast cancer*. Diabetes Care, 2010. **33**(6): p. 1304-8.
35. Armstrong, R.G., J. West, and T.R. Card, *Risk of cancer in inflammatory bowel disease treated with azathioprine: a UK population-based case-control study*. American Journal of Gastroenterology, 2010. **105**(7): p. 1604-9.
36. Currie, C.J., C.D. Poole, and E.A.M. Gale, *The influence of glucose-lowering therapies on cancer risk in type 2 diabetes*. Diabetologia, 2009. **52**(9): p. 1766-77.
37. Schneider, C., S.S. Jick, and C.R. Meier, *Risk of gynecological cancers in users of estradiol/dydrogesterone or other HRT preparations*. Climacteric, 2009. **12**(6): p. 514-24.
38. Haynes, K., et al., *Cancer incidence in The Health Improvement Network*. Pharmacoepidemiology & Drug Safety, 2009. **18**(8): p. 730-6.
39. Simon, T.A., et al., *Malignancies in the rheumatoid arthritis abatacept clinical development programme: An epidemiological assessment*. Annals of the Rheumatic Diseases, 2009. **68**(12): p. 1819-1826.
40. van Staa, T.P. and J.M. Sprafka, *Study of adverse outcomes in women using testosterone therapy*. Maturitas, 2009. **62**(1): p. 76-80.
41. Brauchli, Y.B., et al., *Psoriasis and risk of incident cancer: an inception cohort study with a nested case-control analysis*. Journal of Investigative Dermatology, 2009. **129**(11): p. 2604-12.



42. Smeeth, L., et al., *Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials*. British Journal of Clinical Pharmacology, 2009. **67**(1): p. 99-109.
43. Opatrny, L., et al., *Hormone replacement therapy use and variations in the risk of breast cancer*. BJOG, 2008. **115**(2): p. 169-75; discussion 175.
44. Weiner, M.G., et al., *Hormone therapy and coronary heart disease in young women*. Menopause, 2008. **15**(1): p. 86-93.
45. Lewis, N.R., et al., *No increase in risk of fracture, malignancy or mortality in dermatitis herpetiformis: A cohort study*. Alimentary Pharmacology and Therapeutics, 2008. **27**(11): p. 1140-1147.
46. van Staa, T.P., et al., *Individualizing the risks and benefits of postmenopausal hormone therapy*. Menopause, 2008. **15**(2): p. 374-81.
47. Yang, Y.X., et al., *Chronic statin therapy and the risk of colorectal cancer*. Pharmacoepidemiology and Drug Safety, 2008. **17**(9): p. 869-876.
48. Lewis, J.D., et al., *Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research*. Pharmacoepidemiology & Drug Safety, 2007. **16**(4): p. 393-401.
49. Parker, C., et al., *Rectal and postmenopausal bleeding: Consultation and referral of patients with and without severe mental health problems*. British Journal of General Practice, 2007. **57**(538): p. 371-376.
50. Jones, R., et al., *Alarm symptoms in early diagnosis of cancer in primary care: cohort study using General Practice Research Database*. BMJ, 2007. **334**(7602): p. 1040.
51. Srinivasan, R., et al., *Risk of colorectal cancer in women with a prior diagnosis of gynecologic malignancy*. Journal of Clinical Gastroenterology, 2007. **41**(3): p. 291-296.
52. Jackson, H., et al., *Influence of ursodeoxycholic acid on the mortality and malignancy associated with primary biliary cirrhosis: A population-based cohort study*. Hepatology, 2007. **46**(4): p. 1131-1137.
53. Le Jeune, I., et al., *The incidence of cancer in patients with idiopathic pulmonary fibrosis and sarcoidosis in the UK*. Respiratory Medicine, 2007. **101**(12): p. 2534-40.
54. Hippisley-Cox, J., et al., *Risk of malignancy in patients with schizophrenia or bipolar disorder: Nested case-control study*. Archives of General Psychiatry, 2007. **64**(12): p. 1368-1376.
55. Tannen, R.L., et al., *A simulation using data from a primary care practice database closely replicated the women's health initiative trial*. Journal of Clinical Epidemiology, 2007. **60**(7): p. 686-695.

56. Tannen, R.L., et al., *Estrogen affects post-menopausal women differently than estrogen plus progestin replacement therapy*. Human Reproduction, 2007. **22**(6): p. 1769-77.
57. Vinogradova, Y., et al., *Risk of Colorectal Cancer in Patients Prescribed Statins, Nonsteroidal Anti-Inflammatory Drugs, and Cyclooxygenase-2 Inhibitors: Nested Case-Control Study*. Gastroenterology, 2007. **133**(2): p. 393-402.
58. Yang, Y.-X., et al., *Chronic proton pump inhibitor therapy and the risk of colorectal cancer*. Gastroenterology, 2007. **133**(3): p. 748-54.
59. Gonzalez-Perez, A., et al., *Cancer incidence in a general population of asthma patients*. Pharmacoepidemiology & Drug Safety, 2006. **15**(2): p. 131-8.
60. Gonzalez-Perez, A. and L.A. Garcia Rodriguez, *Breast cancer risk among users of antidepressant medications*. Epidemiology, 2005. **16**(1): p. 101-105.
61. Gonzalez-Perez, A. and L.A. Garcia Rodriguez, *Prostate cancer risk among men with diabetes mellitus (Spain)*. Cancer Causes & Control, 2005. **16**(9): p. 1055-8.
62. Bradbury, B.D., J.B. Wilk, and J.A. Kaye, *Obesity and the risk of prostate cancer (United States)*. Cancer Causes and Control, 2005. **16**(6): p. 637-641.
63. Kaye, J.A. and H. Jick, *Antibiotics and the risk of breast cancer*. Epidemiology, 2005. **16**(5): p. 688-90.
64. Lewis, J.D., et al., *The relationship between time since registration and measured incidence rates in the General Practice Research Database*. Pharmacoepidemiology & Drug Safety, 2005. **14**(7): p. 443-51.
65. Garcia Rodriguez, L.A. and A. Gonzalez-Perez, *Use of antibiotics and risk of breast cancer*. American Journal of Epidemiology, 2005. **161**(7): p. 616-9.
66. Shao, T. and Y.X. Yang, *Cholecystectomy and the risk of colorectal cancer*. American Journal of Gastroenterology, 2005. **100**(8): p. 1813-1820.
67. Van Staa, T.P., et al., *5-Aminosalicylate use and colorectal cancer risk in inflammatory bowel disease: A large epidemiological study*. Gut, 2005. **54**(11): p. 1573-1578.
68. Yang, Y.-X., S. Hennessy, and J.D. Lewis, *Type 2 diabetes mellitus and the risk of colorectal cancer*. Clinical Gastroenterology & Hepatology, 2005. **3**(6): p. 587-94.
69. Gonzalez-Perez, A., G. Ronquist, and L.A. Garcia Rodriguez, *Breast cancer incidence and use of antihypertensive medication in women*. Pharmacoepidemiology & Drug Safety, 2004. **13**(8): p. 581-5.
70. Ronquist, G., et al., *Association between captopril, other antihypertensive drugs and risk of prostate cancer*. Prostate, 2004. **58**(1): p. 50-6.

71. Kaye, J.A. and H. Jick, *Statin use and cancer risk in the General Practice Research Database*. British Journal of Cancer, 2004. **90**(3): p. 635-637.
72. West, J., et al., *Malignancy and mortality in people with coeliac disease: population based cohort study*. BMJ, 2004. **329**(7468): p. 716-9.
73. Garcia Rodriguez, L.A. and A. Gonzalez-Perez, *Risk of breast cancer among users of aspirin and other anti-inflammatory drugs*. British Journal of Cancer, 2004. **91**(3): p. 525-529.
74. Garcia Rodriguez, L.A. and A. Gonzalez-Perez, *Inverse association between nonsteroidal anti-inflammatory drugs and prostate cancer*. Cancer Epidemiology, Biomarkers & Prevention, 2004. **13**(4): p. 649-53.
75. Solaymani-Dodaran, M., et al., *Risk of extra-oesophageal malignancies and colorectal cancer in Barrett's oesophagus and gastro-oesophageal reflux*. Scandinavian Journal of Gastroenterology, 2004. **39**(7): p. 680-5.
76. Yang, Y.X., S. Hennessy, and J.D. Lewis, *Insulin therapy and colorectal cancer risk among type 2 diabetes mellitus patients*. Gastroenterology, 2004. **127**(4): p. 1044-1050.
77. Meier, C.R., S. Schmitz, and H. Jick, *Association between acetaminophen or nonsteroidal antiinflammatory drugs and risk of developing ovarian, breast, or colon cancer*. Pharmacotherapy: The Journal of Human Pharmacology & Drug Therapy, 2002. **22**(3): p. 303-9.
78. Kaye, J.A., et al., *Statin use, hyperlipidaemia, and the risk of breast cancer*. British Journal of Cancer, 2002. **86**(9): p. 1436-9.
79. Garcia-Rodriguez, L.A. and C. Huerta-Alvarez, *Reduced risk of colorectal cancer among long-term users of aspirin and nonaspirin nonsteroidal antiinflammatory drugs*. Epidemiology, 2001. **12**(1): p. 88-93.
80. Meier, C.R., et al., *Angiotensin-converting enzyme inhibitors, calcium channel blockers, and breast cancer*. Archives of Internal Medicine, 2000. **160**(3): p. 349-353.
81. Kaye, J.A., et al., *The incidence of breast cancer in the General Practice Research Database compared with national cancer registration data*. British Journal of Cancer, 2000. **83**(11): p. 1556-8.
82. Langman, M.J., et al., *Effect of anti-inflammatory drugs on overall risk of common cancer: case-control study in general practice research database*. BMJ, 2000. **320**(7250): p. 1642-6.
83. Garcia Rodriguez, L.A. and C. Huerta-Alvarez, *Reduced incidence of colorectal adenoma among long-term users of nonsteroidal antiinflammatory drugs: a pooled analysis of published studies and a new population-based study*. Epidemiology, 2000. **11**(4): p. 376-81.

84. Jick, H., et al., *Calcium-channel blockers and risk of cancer*. Lancet, 1997. **349**(9051): p. 525-8.

**Table 8.2: Frequency of studies that included specific cancer related codes**

Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
<b>Breast</b>				
B34..00	MALIGNANT NEOPLASM OF FEMALE BREAST	Malignant Cancer	17	(100)
B340.00	MALIGNANT NEOPLASM OF NIPPLE AND AREOLA OF FEMALE BREAST	Malignant Cancer	17	(100)
B340000	MALIGNANT NEOPLASM OF NIPPLE OF FEMALE BREAST	Malignant Cancer	17	(100)
B340100	MALIGNANT NEOPLASM OF AREOLA OF FEMALE BREAST	Malignant Cancer	17	(100)
B340z00	MALIGNANT NEOPLASM OF NIPPLE OR AREOLA OF FEMALE BREAST NOS	Malignant Cancer	17	(100)
B341.00	MALIGNANT NEOPLASM OF CENTRAL PART OF FEMALE BREAST	Malignant Cancer	17	(100)
B342.00	MALIGNANT NEOPLASM OF UPPER-INNER QUADRANT OF FEMALE BREAST	Malignant Cancer	17	(100)
B343.00	MALIGNANT NEOPLASM OF LOWER-INNER QUADRANT OF FEMALE BREAST	Malignant Cancer	17	(100)
B344.00	MALIGNANT NEOPLASM OF UPPER-OUTER QUADRANT OF FEMALE BREAST	Malignant Cancer	17	(100)
B345.00	MALIGNANT NEOPLASM OF LOWER-OUTER QUADRANT OF FEMALE BREAST	Malignant Cancer	17	(100)
B346.00	MALIGNANT NEOPLASM OF AXILLARY TAIL OF FEMALE BREAST	Malignant Cancer	17	(100)
B347.00	MALIGNANT NEOPLASM, OVERLAPPING LESION OF BREAST	Malignant Cancer	17	(100)
B34y.00	MALIGNANT NEOPLASM OF OTHER SITE OF FEMALE BREAST	Malignant Cancer	17	(100)
B34y000	MALIGNANT NEOPLASM OF ECTOPIC SITE OF FEMALE BREAST	Malignant Cancer	17	(100)
B34z.00	MALIGNANT NEOPLASM OF FEMALE BREAST NOS	Malignant Cancer	17	(100)
B34..11	CA FEMALE BREAST	Malignant Cancer	16	(94)
B34yz00	MALIGNANT NEOPLASM OF OTHER SITE OF FEMALE BREAST NOS	Malignant Cancer	16	(94)
Byu6.00	[X]MALIGNANT NEOPLASM OF BREAST	Malignant Cancer	15	(88)
174 A	NEOPLASM MALIGNANT BREAST	Malignant Cancer	8	(47)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
174 C	CARCINOMA BREAST	Malignant Cancer	8	(47)
174 CI	CARCINOMA BREAST INDURATED	Malignant Cancer	8	(47)
174 DC	ADENOCARCINOMA BREAST	Malignant Cancer	8	(47)
174 DL	ADENOCARCINOMA BREAST ULCERATION	Malignant Cancer	8	(47)
174 AN	MALIGNANT NEOPLASM NIPPLE	Malignant Cancer	6	(35)
B3...00	MALIG NEOP OF BONE, CONNECTIVE TISSUE, SKIN AND BREAST	Malignant Cancer	5	(29)
B3...11	CARCINOMA OF BONE, CONNECTIVE TISSUE, SKIN AND BREAST	Malignant Cancer	5	(29)
B325100	MALIGNANT MELANOMA OF BREAST	Malignant Cancer	5	(29)
B335100	MALIGNANT NEOPLASM OF SKIN OF CHEST, EXCLUDING BREAST	Malignant Cancer	5	(29)
B335200	MALIGNANT NEOPLASM OF SKIN OF BREAST	Malignant Cancer	5	(29)
B35..00	MALIGNANT NEOPLASM OF MALE BREAST	Malignant Cancer	5	(29)
B350.00	MALIGNANT NEOPLASM OF NIPPLE AND AREOLA OF MALE BREAST	Malignant Cancer	5	(29)
B350000	MALIGNANT NEOPLASM OF NIPPLE OF MALE BREAST	Malignant Cancer	5	(29)
B350100	MALIGNANT NEOPLASM OF AREOLA OF MALE BREAST	Malignant Cancer	5	(29)
B35z.00	MALIGNANT NEOPLASM OF OTHER SITE OF MALE BREAST	Malignant Cancer	5	(29)
B35z000	MALIGNANT NEOPLASM OF ECTOPIC SITE OF MALE BREAST	Malignant Cancer	5	(29)
B35zz00	MALIGNANT NEOPLASM OF MALE BREAST NOS	Malignant Cancer	5	(29)
B3y..00	MALIG NEOP OF BONE, CONNECTIVE TISSUE, SKIN AND BREAST OS	Malignant Cancer	5	(29)
B3z..00	MALIG NEOP OF BONE, CONNECTIVE TISSUE, SKIN AND BREAST NOS	Malignant Cancer	5	(29)
B544.00	MALIGNANT NEOPLASM OF CAROTID BODY	Malignant Cancer	4	(24)
174 PB	PAGET'S DISEASE BREAST	Malignant Cancer	3	(18)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
174 PN	PAGET'S DISEASE NIPPLE	Malignant Cancer	2	(12)
B350z00	MALIGNANT NEOPLASM OF NIPPLE OR AREOLA OF MALE BREAST NOS	Malignant Cancer	2	(12)
B83..00	CARCINOMA IN SITU OF BREAST AND GENITOURINARY SYSTEM	In-Situ	11	(65)
B830.00	CARCINOMA IN SITU OF BREAST	In-Situ	11	(65)
B830000	LOBULAR CARCINOMA IN SITU OF BREAST	In-Situ	11	(65)
B830100	INTRADUCTAL CARCINOMA IN SITU OF BREAST	In-Situ	11	(65)
ByuFG00	[X]OTHER CARCINOMA IN SITU OF BREAST	In-Situ	10	(59)
BB90.00	[M]INTRADUCTAL CARCINOMA, NONINFILTRATING NOS	In-Situ	6	(35)
B825000	CARCINOMA IN SITU OF SKIN OF BREAST	In-Situ	5	(29)
BB9E000	[M]INTRADUCTAL CARCINOMA AND LOBULAR CARCINOMA IN SITU	In-Situ	5	(29)
BB9E.00	[M]LOBULAR CARCINOMA IN SITU	In-Situ	2	(12)
BB92.00	[M] COMEDOCARCINOMA, NON-INFILTRATING	In-Situ	1	(6)
BB96.00	[M]NONINFILTRATING INTRADUCTAL PAPILLARY ADENOCARCINOMA	In-Situ	1	(6)
BB9K.00	[M]PAGET'S DISEASE AND INFILTRATING BREAST DUCT CARCINOMA	Cancer Morphology	13	(76)
BB91100	[M]INFILTRATING DUCT AND LOBULAR CARCINOMA	Cancer Morphology	12	(71)
BB94.00	[M]JUVENILE BREAST CARCINOMA	Cancer Morphology	12	(71)
BB94.11	[M]SECRETORY BREAST CARCINOMA	Cancer Morphology	12	(71)
BB9K000	[M]PAGET'S DISEASE AND INTRADUCTAL CARCINOMA OF BREAST	Cancer Morphology	12	(71)
BB91.00	[M]INFILTRATING DUCT CARCINOMA	Cancer Morphology	11	(65)
BB9F.00	[M]LOBULAR CARCINOMA NOS	Cancer Morphology	11	(65)
BB9G.00	[M]INFILTRATING DUCTULAR CARCINOMA	Cancer Morphology	11	(65)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
BB91000	[M]INTRADUCTAL PAPILLARY ADENOCARCINOMA WITH INVASION	Cancer Morphology	10	(59)
BB93.00	[M] COMEDOCARCINOMA NOS	Cancer Morphology	10	(59)
BB9H.00	[M]INFLAMMATORY CARCINOMA	Cancer Morphology	10	(59)
BB9J.11	[M]PAGET'S DISEASE, BREAST	Cancer Morphology	9	(53)
BBM9.00	[M]CYSTOSARCOMA PHYLLODES, MALIGNANT	Cancer Morphology	9	(53)
BB9J.00	[M]PAGET'S DISEASE, MAMMARY	Cancer Morphology	8	(47)
BB9D.00	[M]MEDULLARY CARCINOMA WITH LYMPHOID STROMA	Cancer Morphology	6	(35)
BB9B.00	[M] MEDULLARY CARCINOMA NOS	Cancer Morphology	2	(12)
BB91.11	[M] DUCT CARCINOMA NOS	Cancer Morphology	1	(6)
BB9B.11	[M] C CELL CARCINOMA	Cancer Morphology	1	(6)
BB9C.00	[M] MEDULLARY CARCINOMA WITH AMYLOID STROMA	Cancer Morphology	1	(6)
BB9L.00	[M] PAGET	Cancer Morphology	1	(6)
BB9M.00	[M] INTRACYSTIC CARCINOMA NOS	Cancer Morphology	1	(6)
B58y000	SECONDARY MALIGNANT NEOPLASM OF BREAST	H/0 & Secondary	13	(76)
B582600	SECONDARY MALIGNANT NEOPLASM OF SKIN OF BREAST	H/0 & Secondary	5	(29)
BB85111	[M]KRUKENBERG TUMOUR	H/0 & Secondary	4	(24)
ZV10300	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF BREAST	H/0	4	(24)
ZV13A00	[V]PERSONAL HISTORY OF NON-NEOPLASTIC BREAST DISEASE	H/0	2	(12)
BA03.00	NEOPLASM OF UNSPECIFIED NATURE OF BREAST	Borderline	6	(35)
B933.00	NEOPLASM OF UNCERTAIN BEHAVIOUR OF BREAST	Borderline	2	(12)
4KJ0.00	*OESTROGEN RECEPTOR POSITIVE TUMOUR (ADDED 1 <sup>ST</sup> APR 2008)	Borderline	1	(6)

[Table 8.2 continued over]



[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
4KJ1.00	*PROGESTERONE RECEPTOR POSITIVE TUMOUR (ADDED 1 <sup>ST</sup> APR 2008)	Borderline	1	(6)
4KJ2.00	*OESTROGEN RECEPTOR NEGATIVE TUMOUR (ADDED 1 <sup>ST</sup> APR 2008)	Borderline	1	(6)
4KJ3.00	*PROGESTERONE RECEPTOR NEGATIVE TUMOUR (ADDED 1 <sup>ST</sup> APR 2008)	Borderline	1	(6)
BB9..00	[M] DUCTAL, LOBULAR, AND MEDULLARY NEOPLASMS	Borderline	1	(6)
BB9z.00	[M] DUCTAL, LOBULAR, OR MEDULLARY NEOPLASM NOS	Borderline	1	(6)
6862100	BREAST NEOPLASM SCREEN ABNORM	Suspected	2	(12)
1J0I.00	SUSPECTED BREAST CANCER	Suspected	1	(6)
8Hn2.00	*FAST TRACK REFERRAL FOR SUSPECTED BREAST CANCER (ADDED 1 <sup>ST</sup> APR 2007)	Suspected	1	(6)
9Np2.00	*SEEN IN FAST TRACK SUSPECTED BREAST CANCER CLINIC (ADDED 1 <sup>ST</sup> APR 2009)	Suspected	1	(6)
217	TUMOUR BREAST BENIGN	Benign	1	(6)
217 AF	FIBROADENOMA BREAST	Benign	1	(6)
B765100	BENIGN NEOPLASM OF SKIN OF BREAST	Benign	1	(6)
B77..00	BENIGN NEOPLASM OF BREAST	Benign	1	(6)
B770.00	BENIGN NEOPLASM OF FEMALE BREAST	Benign	1	(6)
B771.00	BENIGN NEOPLASM OF MALE BREAST	Benign	1	(6)
B77z.00	BENIGN NEOPLASM OF BREAST NOS	Benign	1	(6)
122B.00	NO FH: BREAST CARCINOMA	Not Cancer	1	(6)
1243.11	FH: BREAST CANCER	Not Cancer	1	(6)
585C.00	US SCAN OF BREAST	Not Cancer	1	(6)
610 AC	NONNEOPLASTIC BREAST CONDITION	Not Cancer	1	(6)
610 AD	BREAST NON-NEOPLASTIC DISEASE	Not Cancer	1	(6)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
6862	BREAST NEOPLASM SCREEN	Not Cancer	1	(6)
6862000	BREAST NEOPLASM SCREEN NORMAL	Not Cancer	1	(6)
6862200	BREAST NEOPLASM SCREEN NOS	Not Cancer	1	(6)
7135000	PERCUTANEOUS BIOPSY OF BREAST LESION	Not Cancer	1	(6)
7G26100	INSERTION SKIN EXPANDER INTO SUBCUTANEOUS TISSUE OF BREAST	Not Cancer	1	(6)
7G27200	REMOVAL OF SKIN EXPANDER FROM SUBCUTANEOUS TISSUE OF BREAST	Not Cancer	1	(6)
F1740C	FAMILY HISTORY OF BREAST CARCINOMA	Not Cancer	1	(6)
M002100	CARBUNCLE OF BREAST	Not Cancer	1	(6)
Q007.00	FETUS/NEONATE AFFECTED BY POISON TRANSFERRED PLACENTA/BREAST	Not Cancer	1	(6)
Q007000	FETUS/NEONATE AFFECTED-PLACENTAL/BREAST TRANSFER UNSP POISON	Not Cancer	1	(6)
Q007100	FETUS/NEONATE AFFECTED BY PLACENTAL/BREAST TRANSFER ALCOHOL	Not Cancer	1	(6)
Q007200	FETUS/NEONATE AFFECTED BY PLACENTAL/BREAST TRANSFER NARCOTIC	Not Cancer	1	(6)
Q007300	FETUS/NEONATE AFFECTED-PLACENTA/BREAST TRANSFER HALLUCINOGEN	Not Cancer	1	(6)
Q007400	FETUS/NEONATE AFFECTED-PLACENTAL/BREAST TRANSFER ANTI-INFECT	Not Cancer	1	(6)
Q007411	FETUS/NEONATE AFFECTED-PLACENTAL/BREAST TRANSFER ANTIBIOTIC	Not Cancer	1	(6)
Q007500	FETUS/NEONATE AFFECTED-PLACENTAL/BREAST TRANSFER IMMUNE SERA	Not Cancer	1	(6)
Q007600	FETUS/NEONATE AFFECTED-PLAC./BREAST TRANSFER ANTICONVULSANT	Not Cancer	1	(6)
Q007700	FETUS/NEONATE AFFECTED-PLAC./BREAST TRANSFER ANTICOAGULANT	Not Cancer	1	(6)
Q007900	FETUS/NEONATE AFFECTED-PLAC./BREAST TRANSFER UTERINE DEPRESS	Not Cancer	1	(6)
Q007A00	FETUS/NEONATE AFFECT-PLAC./BREAST TRANSF HYPOGLYCAEMIC AGENT	Not Cancer	1	(6)
Q007B00	FETUS/NEONATE AFFECTED-PLAC./BREAST TRANSFER ENDOCRINE AGENT	Not Cancer	1	(6)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
Q007C00	FETUS/NEONATE AFFECTED-PLAC./BREAST TRANSFER ADDICTIVE DRUG	Not Cancer	1	(6)
Q007w00	FETUS/NEONATE AFFECTED-PLACENTA/BREAST TRANSFER MEDICINE NEC	Not Cancer	1	(6)
Q007x00	FETUS/NEONATE AFFECTED - PLACENTAL/BREAST TRANSFER TOXIC NEC	Not Cancer	1	(6)
Q007y00	FETUS/NEONATE AFFECTED - POISON TRANSFER PLACENTA/BREAST OS	Not Cancer	1	(6)
Q007z00	FETUS/NEONATE AFFECTED - POISON TRANSFER PLACENTA/BREAST NOS	Not Cancer	1	(6)
Q433200	BREAST FEEDING INHIBITORS CAUSING NEONATAL JAUNDICE	Not Cancer	1	(6)
Q433A00	NEONATAL JAUNDICE FROM BREAST MILK INHIBITOR	Not Cancer	1	(6)
Z16..00	BREAST CARE PROCEDURE	Not Cancer	1	(6)
Z174B00	BREAST CARE	Not Cancer	1	(6)
ZL1G100	UNDER CARE OF BREAST SURGEON	Not Cancer	1	(6)
ZL22100	UNDER CARE OF BREAST CARE NURSE	Not Cancer	1	(6)
ZL62100	REFERRAL TO BREAST CARE NURSE	Not Cancer	1	(6)
ZLA2100	SEEN BY BREAST CARE NURSE	Not Cancer	1	(6)
ZV16300	[V]FAMILY HISTORY OF MALIGNANT NEOPLASM OF BREAST	Not Cancer	1	(6)
ZV6C100	[V]FOLLOW-UP CARE INVOLVING PLASTIC SURGERY OF BREAST	Not Cancer	1	(6)
ZV76100	[V]SCREENING FOR MALIGNANT NEOPLASM OF BREAST	Not Cancer	1	(6)
<b>Colorectal</b>				
B13..00	MALIGNANT NEOPLASM OF COLON	Malignant Cancer	23	(100)
B130.00	MALIGNANT NEOPLASM OF HEPATIC FLEXURE OF COLON	Malignant Cancer	23	(100)
B131.00	MALIGNANT NEOPLASM OF TRANSVERSE COLON	Malignant Cancer	23	(100)
B132.00	MALIGNANT NEOPLASM OF DESCENDING COLON	Malignant Cancer	23	(100)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
B133.00	MALIGNANT NEOPLASM OF SIGMOID COLON	Malignant Cancer	23	(100)
B136.00	MALIGNANT NEOPLASM OF ASCENDING COLON	Malignant Cancer	23	(100)
B137.00	MALIGNANT NEOPLASM OF SPLENIC FLEXURE OF COLON	Malignant Cancer	23	(100)
B138.00	MALIGNANT NEOPLASM, OVERLAPPING LESION OF COLON	Malignant Cancer	23	(100)
B13y.00	MALIGNANT NEOPLASM OF OTHER SPECIFIED SITES OF COLON	Malignant Cancer	23	(100)
B13z.00	MALIGNANT NEOPLASM OF COLON NOS	Malignant Cancer	23	(100)
B13z.11	COLONIC CANCER	Malignant Cancer	23	(100)
B14..00	MALIGNANT NEOPLASM OF RECTUM, RECTOSIGMOID JUNCTION AND ANUS	Malignant Cancer	23	(100)
B140.00	MALIGNANT NEOPLASM OF RECTOSIGMOID JUNCTION	Malignant Cancer	23	(100)
B141.00	MALIGNANT NEOPLASM OF RECTUM	Malignant Cancer	23	(100)
B141.11	CARCINOMA OF RECTUM	Malignant Cancer	23	(100)
B141.12	RECTAL CARCINOMA	Malignant Cancer	23	(100)
B14y.00	MALIG NEOP OTHER SITE RECTUM, RECTOSIGMOID JUNCTION AND ANUS	Malignant Cancer	23	(100)
B14z.00	MALIGNANT NEOPLASM RECTUM,RECTOSIGMOID JUNCTION AND ANUS NOS	Malignant Cancer	23	(100)
B134.00	MALIGNANT NEOPLASM OF CAECUM	Malignant Cancer	21	(91)
B134.11	CARCINOMA OF CAECUM	Malignant Cancer	21	(91)
B135.00	MALIGNANT NEOPLASM OF APPENDIX	Malignant Cancer	21	(91)
B142.00	MALIGNANT NEOPLASM OF ANAL CANAL	Malignant Cancer	19	(83)
B142.11	ANAL CARCINOMA	Malignant Cancer	19	(83)
B142000	MALIGNANT NEOPLASM OF CLOACOGENIC ZONE	Malignant Cancer	19	(83)
B143.00	MALIGNANT NEOPLASM OF ANUS UNSPECIFIED	Malignant Cancer	19	(83)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
B18y200	MALIGNANT NEOPLASM OF MESORECTUM	Malignant Cancer	17	(74)
B1z0.11	CANCER OF BOWEL	Malignant Cancer	17	(74)
1530AD	ADENOCARCINOMA ASCENDING COLON	Malignant Cancer	10	(43)
1533AD	ADENOCARCINOMA SIGMOID COLON	Malignant Cancer	10	(43)
1538AD	ADENOCARCINOMA COLON	Malignant Cancer	10	(43)
1538AN	MALIGNANT NEOPLASM LARGE BOWEL NONRECTAL	Malignant Cancer	10	(43)
1538C	COLON CARCINOMA	Malignant Cancer	10	(43)
1538CN	LARGE BOWEL CARCINOMA NONRECTAL	Malignant Cancer	10	(43)
1539A	MALIGNANT NEOPLASM BOWEL	Malignant Cancer	10	(43)
1541A	MALIGNANT NEOPLASM RECTUM	Malignant Cancer	10	(43)
1541C	RECTUM CARCINOMA	Malignant Cancer	10	(43)
1539C	CARCINOMA BOWEL	Malignant Cancer	9	(39)
1530AC	MALIGNANT NEOPLASM CAECUM	Malignant Cancer	8	(35)
1530CC	CARCINOMA CAECUM	Malignant Cancer	8	(35)
1533A	MALIGNANT NEOPLASM SIGMOID	Malignant Cancer	8	(35)
1538B	SARCOMA COLON	Malignant Cancer	5	(22)
1538A	MALIGNANT NEOPLASM LARGE INTESTINE	Malignant Cancer	4	(17)
1539AT	MALIGNANT NEOPLASM INTESTINE	Malignant Cancer	3	(13)
B18y000	MALIGNANT NEOPLASM OF MESOCOLON	Malignant Cancer	3	(13)
1529A	MALIGNANT NEOPLASM SMALL INTESTINE	Malignant Cancer	2	(9)
1529C	CARCINOMA SMALL INTESTINE	Malignant Cancer	2	(9)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
1542A	MALIGNANT NEOPLASM ANAL CANAL	Malignant Cancer	2	(9)
1542C	CARCINOMA ANAL CANAL	Malignant Cancer	2	(9)
1736AN	MALIGNANT NEOPLASM ANUS	Malignant Cancer	2	(9)
1736CN	CARCINOMA ANUS	Malignant Cancer	2	(9)
4M1..00	DUKES STAGING SYSTEM	Malignant Cancer	2	(9)
4M10.00	DUKES STAGE A	Malignant Cancer	2	(9)
4M11.00	DUKES STAGE B	Malignant Cancer	2	(9)
4M12.00	DUKES STAGE C1	Malignant Cancer	2	(9)
4M13.00	DUKES STAGE C2	Malignant Cancer	2	(9)
4M14.00	DUKES STAGE D	Malignant Cancer	2	(9)
9Ow1.00	*BOWEL CANCER DETECTED BY NATIONAL SCREENING PROGRAMME (ADDED 1 <sup>ST</sup> OCT 2007)	Malignant Cancer	2	(9)
B12..00	MALIGNANT NEOPLASM OF SMALL INTESTINE AND DUODENUM	Malignant Cancer	2	(9)
B124.00	MALIGNANT NEOPLASM, OVERLAPPING LESION OF SMALL INTESTINE	Malignant Cancer	2	(9)
B12y.00	MALIGNANT NEOPLASM OF OTHER SPECIFIED SITE SMALL INTESTINE	Malignant Cancer	2	(9)
B12z.00	MALIGNANT NEOPLASM OF SMALL INTESTINE NOS	Malignant Cancer	2	(9)
B139.00	*HEREDITARY NONPOLYPOSIS COLON CANCER (ADDED 1 <sup>ST</sup> OCT 2010)	Malignant Cancer	2	(9)
B180200	MALIGNANT NEOPLASM OF RETROCAECAL TISSUE	Malignant Cancer	1	(4)
B18y100	MALIGNANT NEOPLASM OF MESOCAECUM	Malignant Cancer	1	(4)
B803.00	CARCINOMA IN SITU OF COLON	In-Situ	14	(61)
B803000	CARCINOMA IN SITU OF HEPATIC FLEXURE OF COLON	In-Situ	14	(61)
B803100	CARCINOMA IN SITU OF TRANSVERSE COLON	In-Situ	14	(61)
[Appendix A.3 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
B803200	CARCINOMA IN SITU OF DESCENDING COLON	In-Situ	14	(61)
B803300	CARCINOMA IN SITU OF SIGMOID COLON	In-Situ	14	(61)
B803600	CARCINOMA IN SITU OF ASCENDING COLON	In-Situ	14	(61)
B803700	CARCINOMA IN SITU OF SPLENIC FLEXURE OF COLON	In-Situ	14	(61)
B803z00	CARCINOMA IN SITU OF COLON NOS	In-Situ	14	(61)
B804.00	CARCINOMA IN SITU OF RECTUM AND RECTOSIGMOID JUNCTION	In-Situ	14	(61)
B804100	CARCINOMA IN SITU OF RECTUM	In-Situ	14	(61)
B804z00	CARCINOMA IN SITU OF RECTUM OR RECTOSIGMOID JUNCTION NOS	In-Situ	14	(61)
B804000	CARCINOMA IN SITU OF RECTOSIGMOID JUNCTION	In-Situ	13	(57)
B803400	CARCINOMA IN SITU OF CAECUM	In-Situ	8	(35)
B803500	CARCINOMA IN SITU OF APPENDIX	In-Situ	3	(13)
B805.00	CARCINOMA IN SITU OF ANAL CANAL	In-Situ	3	(13)
B806.00	CARCINOMA IN SITU OF ANUS NOS	In-Situ	3	(13)
B807.00	CARCINOMA IN SITU OF OTHER AND UNSPECIFIED SMALL INTESTINE	In-Situ	2	(9)
B807z00	CARCINOMA IN SITU OTHER AND UNSPECIFIED SMALL INTESTINE NOS	In-Situ	2	(9)
ByuF000	[X]CARCINOMA IN SITU/OTHER+UNSPECIFIED PARTS OF INTESTINE	In-Situ	2	(9)
B805000	*ANAL INTRAEPITHELIAL NEOPLASIA GRADE III (ADDED 1 <sup>ST</sup> MAR 2003)	In-Situ	1	(4)
BB5R600	[M]MUCOCARCINOID TUMOUR, MALIGNANT	Cancer Morphology	10	(43)
BB5N100	[M]ADENOCARCINOMA IN ADENOMATOUS POLPOSIS COLI	Cancer Morphology	5	(22)
BB48.00	[M]BASALOID CARCINOMA	Cancer Morphology	1	(4)
BB5L100	[M]ADENOCARCINOMA IN ADENOMATOUS POLYP	Cancer Morphology	1	(4)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
BB5L300	[M]ADENOCARCINOMA IN MULTIPLE ADENOMATOUS POLYPS	Cancer Morphology	1	(4)
BB82111	[M]COLLOID ADENOCARCINOMA	Cancer Morphology	1	(4)
B575.00	SECONDARY MALIGNANT NEOPLASM OF LARGE INTESTINE AND RECTUM	H/0 & Secondary	19	(83)
B575z00	SECONDARY MALIG NEOP OF LARGE INTESTINE OR RECTUM NOS	H/0 & Secondary	18	(78)
B575000	SECONDARY MALIGNANT NEOPLASM OF COLON	H/0 & Secondary	17	(74)
B575100	SECONDARY MALIGNANT NEOPLASM OF RECTUM	H/0 & Secondary	16	(70)
B574.00	SECONDARY MALIGNANT NEOPLASM OF SMALL INTESTINE AND DUODENUM	H/0 & Secondary	2	(9)
B574z00	SECONDARY MALIG NEOP OF SMALL INTESTINE OR DUODENUM NOS	H/0 & Secondary	2	(9)
ZV10017	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF RECTUM	H/0	3	(13)
ZV10013	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF INTESTINE	H/0	2	(9)
ZV10014	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF LARGE INTESTINE	H/0	2	(9)
ZV10011	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF ANUS	H/0	1	(4)
BB5N.00	[M]ADENOMATOUS AND ADENOCARCINOMATOUS POLYPS OF COLON	Borderline	10	(43)
BB5Nz00	[M]ADENOMATOUS OR ADENOCARCINOMATOUS POLYPS OF THE COLON NOS	Borderline	10	(43)
B902400	NEOPLASM OF UNCERTAIN BEHAVIOUR OF COLON	Borderline	7	(30)
B902500	NEOPLASM OF UNCERTAIN BEHAVIOUR OF RECTUM	Borderline	7	(30)
B902.00	NEOP OF UNCERTAIN BEHAVIOUR STOMACH, INTESTINES AND RECTUM	Borderline	5	(22)
2304	TUMOUR RECTAL	Borderline	4	(17)
B902z00	NEOP OF UNCERTAIN BEHAVIOUR STOMACH, INTESTINE OR RECTUM NOS	Borderline	4	(17)
B905200	NEOPLASM OF UNCERTAIN BEHAVIOUR OF ANAL CANAL AND SPHINCTER	Borderline	4	(17)
B902600	NEOPLASM OF UNCERTAIN OR UNKNOWN BEHAVIOUR OF APPENDIX	Borderline	2	(9)
[Table 8.2 continued over]				



[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
BB5L.00	[M]ADENOMATOUS AND ADENOCARCINOMATOUS POLYPS	Borderline	1	(4)
BB5Lz00	[M]ADENOMATOUS OR ADENOCARCINOMATOUS POLYP NOS	Borderline	1	(4)
BB5N.11	[M]ADENOMA OR OR ADENOCARCINOMA IN POLYPOSIS COLI	Borderline	1	(4)
BB5U.00	[M]VILLOUS ADENOMAS AND ADENOCARCINOMAS	Borderline	1	(4)
8Hn4.00	*FAST TRACK REFERRAL FOR SUSPECTED COLORECTAL CANCER (ADDED 1 <sup>ST</sup> APR 2007)	Suspected	2	(9)
8CAo.00	*PATIENT GIVEN ADVICE ABOUT BOWEL CANCER (ADDED 1 <sup>ST</sup> OCT 2007)	Suspected	1	(4)
9Np7.00	*SEEN IN FAST TRACK SUSPECTED COLORECTAL CANCER CLINIC (ADDED 1 <sup>ST</sup> APR 2009)	Suspected	1	(4)
2114	BENIGN NEOPLASM RECTUM	Benign	3	(13)
2114B	BENIGN NEOPLASM ANORECTAL	Benign	3	(13)
2114C	BENIGN NEOPLASM RECTOSIGMOID JUNCTION	Benign	3	(13)
B713.00	BENIGN NEOPLASM OF COLON	Benign	3	(13)
B713000	BENIGN NEOPLASM OF HEPATIC FLEXURE OF COLON	Benign	3	(13)
B713100	BENIGN NEOPLASM OF TRANSVERSE COLON	Benign	3	(13)
B713200	BENIGN NEOPLASM OF DESCENDING COLON	Benign	3	(13)
B713300	BENIGN NEOPLASM OF SIGMOID COLON	Benign	3	(13)
B713600	BENIGN NEOPLASM OF ASCENDING COLON	Benign	3	(13)
B713700	BENIGN NEOPLASM OF SPLENIC FLEXURE OF COLON	Benign	3	(13)
B713z00	BENIGN NEOPLASM OF COLON NOS	Benign	3	(13)
B714.00	BENIGN NEOPLASM OF RECTUM AND ANAL CANAL	Benign	3	(13)
B714000	BENIGN NEOPLASM OF RECTOSIGMOID JUNCTION	Benign	3	(13)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
B714100	BENIGN NEOPLASM OF RECTUM	Benign	3	(13)
B714z00	BENIGN NEOPLASM OF RECTUM OR ANAL CANAL NOS	Benign	3	(13)
B718300	BENIGN NEOPLASM OF MESOCOLON	Benign	3	(13)
2112N	BENIGN NEOPLASM SMALL INTESTINE	Benign	2	(9)
2113	BENIGN NEOPLASM INTESTINE	Benign	2	(9)
B712.00	BENIGN NEOPLASM OF SMALL INTESTINE AND DUODENUM	Benign	2	(9)
B712z00	BENIGN NEOPLASM OF SMALL INTESTINE OR DUODENUM NOS	Benign	2	(9)
B718400	BENIGN NEOPLASM OF MESORECTUM	Benign	2	(9)
ByuG300	[X]BENIGN NEOPLASM/OTHER+UNSPECIFD PARTS OF SMALL INTESTINE	Benign	2	(9)
2113LC	BENIGN LYMPHOMA COLON	Benign	1	(4)
2113PA	POLYP ADENOMATOUS COLON	Benign	1	(4)
B713400	BENIGN NEOPLASM OF CAECUM	Benign	1	(4)
B713500	BENIGN NEOPLASM OF APPENDIX	Benign	1	(4)
B713800	BENIGN NEOPLASM OF COLOSTOMY SITE	Benign	1	(4)
B713900	BENIGN NEOPLASM OF ILEOCAECAL VALVE	Benign	1	(4)
B714111	BENIGN PAPILLOMA RECTUM	Benign	1	(4)
B714200	BENIGN NEOPLASM OF ANAL CANAL	Benign	1	(4)
B714300	BENIGN NEOPLASM OF ANUS NOS	Benign	1	(4)
BB5L000	[M]ADENOMATOUS POLYP NOS	Benign	1	(4)
BB5L011	[M]POLYPOID ADENOMA	Benign	1	(4)
BB5N000	[M]ADENOMATOUS POLYPOSIS COLI	Benign	1	(4)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
BB5N011	[M]ADENOMATOSIS NOS	Benign	1	(4)
BB5N012	[M]FAMILIAL POLYPOSIS COLI	Benign	1	(4)
BB5N200	[M]MULTIPLE ADENOMATOUS POLYPS	Benign	1	(4)
BB5N211	[M]MULTIPLE POLYPOSIS	Benign	1	(4)
PB2..00	ATRESIA AND STENOSIS OF LARGE INTESTINE/RECTUM/ANAL CANAL	Not Cancer	4	(17)
1241.12	FH: BOWEL CANCER	Not Cancer	2	(9)
6864	LARGE BOWEL NEOPLASM SCREEN	Not Cancer	2	(9)
6864.11	COLON NEOPLASM SCREEN	Not Cancer	2	(9)
6864.12	RECTAL NEOPLASM SCREEN	Not Cancer	2	(9)
7211200	CORRECTION OF EPICANTHUS	Not Cancer	2	(9)
7211300	CORRECTION OF TELECANTHUS, UNSPECIFIED	Not Cancer	2	(9)
7409011	PRIMARY CORRECTION OF ALAR CARTILAGE	Not Cancer	2	(9)
7733300	REANASTOMOSIS RECTUM-ANAL CANAL CORRECT CONG RECTAL ATRESIA	Not Cancer	2	(9)
773C000	LASER RECANALISATION OF BOWEL NEC	Not Cancer	2	(9)
7905000	CORRECT TOTAL ANOMAL PULM VENOUS CONNECT TO SUPRACARD VESSEL	Not Cancer	2	(9)
7A20300	ENDARTERECTOMY AND PATCH REPAIR OF CAROTID ARTERY	Not Cancer	2	(9)
7A20311	CAROTID ENDARTERECTOMY AND PATCH	Not Cancer	2	(9)
7A20400	ENDARTERECTOMY OF CAROTID ARTERY NEC	Not Cancer	2	(9)
7H01011	CORRECTION OF PECTUS CARINATUM	Not Cancer	2	(9)
7L0E200	CENTRALISATION CARPUS- CORRECTN CONGENITAL DEFORMITY FOREARM	Not Cancer	2	(9)
7L1H.12	DIRECT CURRENT CARDIAC SHOCK	Not Cancer	2	(9)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
7L1H000	DIRECT CURRENT CARDIOVERSION	Not Cancer	2	(9)
AB2yz17	CANDIDIASIS RECTUM	Not Cancer	2	(9)
F1539C	FAMILY HISTORY OF BOWEL CANCER	Not Cancer	2	(9)
K0828	ENDARTERECTOMY CAROTID ARTERY	Not Cancer	2	(9)
L244012	RECTOCELE AFFECTING OBSTETRIC CARE	Not Cancer	2	(9)
L244312	RECTOCELE COMPLICATING ANTENATAL CARE - BABY NOT DELIVERED	Not Cancer	2	(9)
L244412	RECTOCELE COMPLICATING POSTPARTUM CARE - BABY DELIVERED PREV	Not Cancer	2	(9)
Q407300	NEONATAL CANDIDIASIS OF INTESTINE	Not Cancer	2	(9)
Z174A00	BOWEL CARE	Not Cancer	2	(9)
Z9...00	INDIRECT CARE PROCEDURES	Not Cancer	2	(9)
ZL1GG00	UNDER CARE OF COLORECTAL SURGEON	Not Cancer	2	(9)
ZV76400	[V]SCREENING FOR MALIGNANT NEOPLASM OF COLON OR RECTUM	Not Cancer	2	(9)
25Q3.00	O/E - PR - RECTAL MASS	Not Cancer	1	(4)
7411600	REMOVAL OF ANTROCHOANAL POLYP	Not Cancer	1	(4)
771G400	COLONOSCOPIC POLYPECTOMY	Not Cancer	1	(4)
7722.11	OPEN OPERATION ON RECTAL POLYP	Not Cancer	1	(4)
7722.12	OPEN POLYPECTOMY OF RECTUM	Not Cancer	1	(4)
7726111	PERANAL EXCISION OF RECTAL POLYP	Not Cancer	1	(4)
7726112	PERANAL POLYPECTOMY OF RECTUM	Not Cancer	1	(4)
7726212	PERANAL DESTRUCTION OF RECTAL POLYP	Not Cancer	1	(4)
7731200	EXCISION OF ANAL POLYP	Not Cancer	1	(4)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
B713.11	COLON POLYP	Not Cancer	1	(4)
H110000	CHOANAL POLYP	Not Cancer	1	(4)
J41y000	PSEUDOPOLYPOSIS OF COLON	Not Cancer	1	(4)
J570.00	ANAL AND RECTAL POLYP	Not Cancer	1	(4)
J570000	ANAL POLYP	Not Cancer	1	(4)
J570100	RECTAL POLYP	Not Cancer	1	(4)
J570z00	ANAL AND RECTAL POLYP NOS	Not Cancer	1	(4)
J578.00	*COLONIC POLYP (ADDED 1 <sup>ST</sup> JAN 2004)	Not Cancer	1	(4)
J578.11	*POLYP OF COLON (ADDED 1 <sup>ST</sup> JAN 2004)	Not Cancer	1	(4)
<b>Prostate</b>				
B46..00	MALIGNANT NEOPLASM OF PROSTATE	Malignant Cancer	11	(100)
4M0..00	GLEASON GRADING OF PROSTATE CANCER	Malignant Cancer	7	(64)
4M00.00	GLEASON PROSTATE GRADE 2-4 (LOW)	Malignant Cancer	6	(55)
4M01.00	GLEASON PROSTATE GRADE 5-7 (MEDIUM)	Malignant Cancer	6	(55)
4M02.00	GLEASON PROSTATE GRADE 8-10 (HIGH)	Malignant Cancer	6	(55)
185 A	MALIGNANT NEOPLASM PROSTATE	Malignant Cancer	4	(36)
185 C	PROSTATE CARCINOMA	Malignant Cancer	4	(36)
185 CA	ADENOCARCINOMA PROSTATE	Malignant Cancer	4	(36)
B834.00	CARCINOMA IN SITU OF PROSTATE	In-Situ	6	(55)
B58y500	SECONDARY MALIGNANT NEOPLASM OF PROSTATE	H/0 & Secondary	8	(73)
ZV10415	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF PROSTATE	H/0	2	(18)
[Table 8.2 continued over]				

[Table 8.2 continued]				
Read/OXMIS Code	Description	Classification	No of studies that included the specific code in their code list (%)	
1427000	*H/O: PROSTATE CANCER (ADDED 1 <sup>ST</sup> APR 2011)	H/O	1	(9)
B915.00	NEOPLASM OF UNCERTAIN BEHAVIOUR OF PROSTATE	Borderline	2	(18)
1J08.00	SUSPECTED PROSTATE CANCER	Suspected	1	(9)

\*Highlighted codes are those that were added to the Read code dictionary during the period the included studies were conducted. The earliest publication year among studies where code lists were provided was 2005, therefore codes added to the dictionary from 2003 have been highlighted (allowing for a 2 year lag period between code list creation and publication).

## Appendix A.1: Chapter 2 systematic review published in the Journal of Pharmacoepidemiology and Drug Safety

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2015; 24: 11–18

Published online 24 November 2014 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/pds.3729

---

### REVIEW

---

## The identification of incident cancers in UK primary care databases: a systematic review<sup>†</sup>

Michael Rañopa<sup>1\*</sup>, Ian Douglas<sup>1</sup>, Tjeerd van Staa<sup>1,2,3</sup>, Liam Smeeth<sup>1</sup>, Olaf Klungel<sup>3</sup>, Robert Reynolds<sup>4</sup> and Krishnan Bhaskaran<sup>1</sup>

<sup>1</sup>Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

<sup>2</sup>Institute of Population Health, University of Manchester, Manchester, United Kingdom

<sup>3</sup>Utrecht University, Faculty of Science, Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht, The Netherlands

<sup>4</sup>Epidemiology, Pfizer Inc., New York, USA

### ABSTRACT

**Purpose** UK primary care databases are frequently used in observational studies with cancer outcomes. We aimed to systematically review methods used by such studies to identify and validate incident cancers of the breast, colorectum, and prostate.

**Methods** Medline and Embase (1980–2013) were searched for UK primary care database studies with incident breast, colorectal, or prostate cancer outcomes. Data on the methods used for case ascertainment were extracted and summarised. Questionnaires were sent to corresponding authors to obtain details about case ascertainment.

**Results** Eighty-four studies of breast ( $n = 51$ ), colorectal ( $n = 54$ ), and prostate cancer ( $n = 31$ ) were identified; 30 examined >1 cancer type. Among the 84 studies, 57 defined cancers using only diagnosis codes, while 27 required further evidence such as chemotherapy. Few studies described methods used to create cancer code lists ( $n = 5$ ); or made lists available directly ( $n = 5$ ). Twenty-eight code lists were received on request from study authors. All included malignant neoplasm diagnosis codes, but there was considerable variation in the specific codes included which was not explained by coding dictionary changes. Code lists also varied in terms of other types of codes included, such as in-situ, cancer morphology, history of cancer, and secondary/suspected/borderline cancer codes.

**Conclusions** In UK primary care database studies, methods for identifying breast, colorectal, and prostate cancers were often unclear. Code lists were often unavailable, and where provided, we observed variation in the individual codes and types of codes included. Clearer reporting of methods and publication of code lists would improve transparency and reproducibility of studies. Copyright © 2014 John Wiley & Sons, Ltd.

**KEY WORDS**—primary care database; cancer; neoplasm; case identification; Read code; pharmacoepidemiology

Received 23 October 2013; Revised 24 September 2014; Accepted 2 October 2014

### INTRODUCTION

UK primary care databases have played an important role in medical research, thanks to their size, representativeness, and the richness of patient data available.<sup>1</sup> Longitudinal data held by the databases include clinical diagnoses, symptoms, prescription therapy, referrals to secondary care, and test results.<sup>1</sup> Studies have frequently utilised UK primary databases such as the Clinical Practice Research Datalink (CPRD),

The Health Improvement Network (THIN), and QRESEARCH to examine drug–cancer associations, disease–cancer relationships, and cancer incidence.<sup>2–4</sup>

In UK primary care settings, Read codes are the standard hierarchical classification system used to record medical information. There are approximately 250 000 Read codes used to record patient diagnoses, symptoms, and processes of care (e.g. referrals to secondary care).<sup>5</sup> The CPRD, THIN, and QRESEARCH currently store all coded medical information according to the Read code system, but the CPRD previously operated using the Oxford Medical Information System (OXMIS).

Identification of cases of disease from these databases is usually a multistage process which begins by compiling a list of selected medical codes, and then matching these codes to patient records.<sup>6</sup> The selection

\*Correspondence to: M. Rañopa, Department of Non-Communicable Disease Epidemiology, LSHTM, Keppel Street, London, United Kingdom. E-mail: Michael.Rañopa@lshtm.ac.uk

<sup>†</sup>Preliminary findings from this study were presented at the International Conference on Pharmacoepidemiology and Therapeutic Risk Management, Montréal, Canada, 2013 (abstract number 609, Pharmacoepi Drug Saf 2013 vol 22 s1 p303).

of medical codes is complex and dependent on several factors such as the study question, clinical background, and experience of the researcher. In a recent CPRD study, low agreement was reported between four medically trained raters when selecting stroke diagnosis codes in a controlled experiment.<sup>7</sup> This lack of agreement may influence findings when studies are investigating similar research questions.<sup>8,9</sup> However, whether this variation occurs among published studies or for other diseases is unknown.

This systematic review aims to investigate current methods used by UK primary care database studies to define, identify, and validate incident cancer cases of the breast, colorectum, and prostate—three of the most common cancers in the UK.<sup>10</sup>

## METHODS

### *Databases and sources*

MEDLINE and EMBASE were searched between Jan 1980 and April 2013 using MeSH terms. Reference lists of relevant studies were also screened for publications that may have been missed by the initial database search. Bibliographies of the CPRD, THIN, QRESEARCH, and the Boston Collaborative Drug Surveillance Program were also screened to identify additional articles that may have been missed by the initial search.<sup>2–4,11</sup>

### *Search keywords and terms*

The search of MEDLINE (8 April 2013) included exploded key terms to identify publications that utilised a UK primary care database and examined incident cancer as an outcome of interest. For EMBASE, which does not use the MeSH classification system, we used the nearest equivalent search terms from the EMBASE indexing system.

### *MEDLINE MeSH terms*

The following MeSH keywords were used in the primary search:

*[Malignant or Cancer or Neoplasm (plus all sub-terms in the MeSH tree)] and [ [GPRD; CPRD; THIN; QRESEARCH; and DIN-LINK (and exploded synonyms)] or [Database (plus all sub-terms in the classification tree)] ]*

### *Inclusion and exclusion criteria*

A publication was considered for initial inclusion when incident cancers of the breast, colorectum, or prostate were included as primary or secondary outcomes, and a UK primary care database was utilised

as a data source. Studies with a main outcome of prevalent, recurring, or metastatic cancer were excluded. Articles presented as conference abstracts, review articles, or letters to the editor were also excluded.

### *Procedure*

Titles and abstracts were initially screened; full-text versions were then obtained and examined to determine whether they met inclusion criteria. Data were extracted from each manuscript including first author, year of publication; study type (e.g. drug safety, epidemiological, or incidence); database(s); cancer outcome(s) of interest; methods used to create code lists (as reported in the paper, e.g. Methods section or supplementary material); case definitions; validation methods and results.

For each study, an electronic copy of the study code list was requested, and the first author was sent a questionnaire which included specific questions on the development of their code list(s). Details of the questionnaire are given in Appendix 1. Three emails were sent to authors: first, the corresponding author was contacted; if no response was received after 3 weeks, then a reminder email was sent to the same author and additionally to the first or last author (if different); a final reminder was sent after a further 3 weeks if necessary. If an error reply was received stating the email address had expired, we searched for a current email address in more recent publications and through an internet search engine.

Medical codes were classified into eight groups: malignant neoplasms; in-situ tumours; malignant morphology; secondary or history of cancer; borderline (uncertain whether malignant or benign); suspected (suspected cancer, abnormal screening test, or fast track referral); benign tumours; and non-cancerous codes (procedure, or condition that was not related to a direct malignant neoplasm diagnosis). Codes were stratified by cancer site and study type. The ICD-10 dictionary and medical references were used to aid in the classification of OXMIS and Read codes.<sup>12–15</sup> All codes were reviewed and classified by KB, LS, and MR, and any disagreements were reviewed again until resolved. All studies in the review were published since the release of the 5-byte Read or Read version 2 dictionary so study code lists were based on the same broad dictionary version; however, codes are continually added to the dictionary over time (though never removed). To assess whether variation in study code lists might have been driven by such changes over time, we obtained a full list of code additions (updates documented 6-monthly from 1991 to 2013) from the NHS Health and Social Care Information Centre and used this to identify codes added during the time period over which the studies



were conducted (which we assumed to be in the 2-year period prior to year of publication).

RESULTS

Databases, cancer site, and study type

Overall, 84 relevant studies were included in this review (Figure 1 & Appendix 2). Studies utilised the CPRD (*n*=63); THIN (*n*=9); QRESEARCH (*n*=10); both the CPRD and THIN (*n*=1); and both the CPRD and QRESEARCH (*n*=1). Of the 84 studies, 30 examined >1 cancer types included in this review: breast (*n*=51); colorectal (*n*=54); and prostate cancer (*n*=31). A broad range of study types were included: 51 examined the association between drug use and cancer; 28 examined cancer incidence among patients with a particular disease or symptom; and 5 estimated population-level cancer incidence (Figure 1).

Study code list creation, availability, and comparison

In total, only 5 of the 84 studies (6%) described methods used to create study code lists (Table 1). Five studies (6%) included details directly in the publication, 2 (2%) included the list itself as an appendix,<sup>16,17</sup> and 3 (4%) by stating which Read code chapters and sections were used; a further 6 (7%) stated that the list was 'available on request'.<sup>18-20</sup>

Overall, there were 43 responses from 84 questionnaires sent to the authors (Figure 1 and Table 1). Thirty-seven (86%) studies reported using a keyword search of cancer related terms to identify potential cancer related codes; 26 (60%) utilised a previous code list; and 43 (100%) consulted with a health care professional during the creation of the study code list. For all studies, >1 assessor reviewed the code list.

In total, 28 of a potential 84 (33%) study code lists were received (Figure 1); frequencies of all codes included across the 28 studies are provided in Appendix

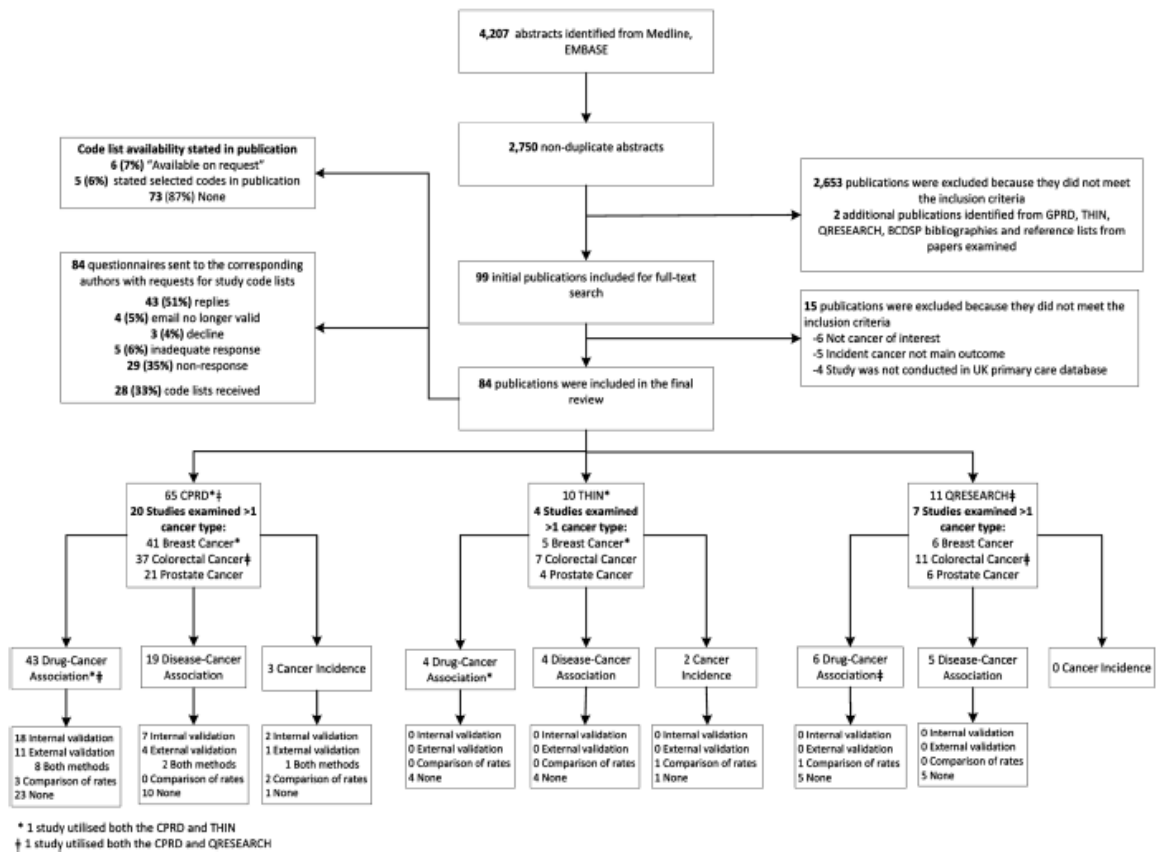


Figure 1. Flow diagram of article search, retrieval, and review process, code list availability and questionnaire replies; database, study type, and validation methods

Table 1. Code list availability, questionnaire replies, and comparison of lists received, by cancer and study type

	All studies	Cancer type*			Study type		
		Breast	Colorectal	Prostate	Drug-cancer	Disease-cancer	Incidence
	<i>n</i> (column %, <i>n/N</i> )						
Total number of studies	<i>N</i> = 84	<i>N</i> = 51	<i>N</i> = 54	<i>N</i> = 31	<i>N</i> = 51	<i>N</i> = 28	<i>N</i> = 5
Any code list creation methods reported <sup>†</sup>	5 (6)	4 (8)	3 (2)	2 (6)	3 (6)	2 (7)	0 (0)
Code list availability in publication							
Available on request	6 (7)	2 (4)	6 (11)	2 (6)	3 (6)	2 (7)	1 (20)
Stated in publication	5 (6)	4 (8)	4 (7)	3 (10)	4 (8)	1 (4)	0 (0)
None	73 (87)	45 (88)	44 (81)	26 (86)	44 (86)	25 (89)	4 (80)
Questionnaire results: number of replies	<i>N</i> = 43	<i>N</i> = 24	<i>N</i> = 30	<i>N</i> = 15	<i>N</i> = 30	<i>N</i> = 10	<i>N</i> = 3
Keyword-synonym search	37 (86)	21 (88)	24 (80)	12 (80)	25 (83)	9 (90)	3 (100)
Utilisation of previous study code list	26 (60)	15 (63)	14 (47)	9 (60)	18 (60)	7 (70)	1 (33)
Consultation with health professional	43 (100)	24 (100)	30 (100)	15 (100)	30 (100)	10 (100)	3 (100)
Number of study code lists obtained	<i>N</i> = 28	<i>N</i> = 17	<i>N</i> = 23	<i>N</i> = 11	<i>N</i> = 21	<i>N</i> = 5	<i>N</i> = 2
Studies including specific code-types:							
Malignant neoplasm	28 (100)	17 (100)	23 (100)	11 (100)	21 (100)	5 (100)	2 (100)
In-situ	20 (71)	12 (70)	15 (65)	6 (55)	13 (62)	5 (100)	2 (100)
Malignant morphology	17 (61)	13 (76)	11 (48)	– (–) <sup>‡</sup>	12 (57)	3 (60)	2 (100)
Secondary or history of cancer	20 (71)	13 (76)	17 (74)	8 (73)	16 (76)	2 (40)	2 (100)
Non-malignant codes	16 (57)	8 (47)	11 (48)	2 (18)	10 (48)	4 (80)	2 (100)
Borderline codes	16 (57)	7 (41)	11 (48)	2 (18)	10 (48)	4 (80)	2 (100)
Suspected	3 (10)	2 (11)	2 (9)	1 (9)	1 (5)	1 (20)	1 (50)
Benign tumour codes	5 (18)	2 (11)	4 (17)	0 (0)	2 (10)	1 (20)	2 (100)
Non-cancerous or site-unrelated	4 (14)	1 (6)	4 (17)	0 (0)	1 (5)	2 (40)	1 (50)

\*One study could contribute to >1 cancer type.

<sup>†</sup>Code list creation methods include: keyword search of dictionary, review of code list by health professional, and utilisation of previous code list.

<sup>‡</sup>There are no malignant morphology codes for prostate cancer found among the 11 prostate cancer studies.

3. All 28 studies included malignant neoplasm diagnosis codes, but there was variation in the specific codes used: for breast cancer, 42 malignant disease codes were included across lists, but only 15 were included by all studies. The variation was not explained by changes in the Read code dictionary: all 42 codes were in the dictionary throughout the period when these studies were conducted (Appendix 3). We found similar variation for colorectal cancer (64 malignant codes mentioned but only 18 appeared in all lists; all but 2 of the 64 codes were in the Read dictionary throughout); and for prostate cancer (8 malignant codes mentioned but only 1 appeared in all lists; all 8 codes present in the Read dictionary throughout). There was also variability between lists in terms of other types of codes included: 20/28 (71%) code lists included in-situ tumours; 17 (61%) included malignant morphology codes; 20 (71%) included secondary or history of cancer codes; 16 (57%) included 'borderline' codes; and 3 (10%) included suspected codes. In addition, a few lists included benign (*n* = 5, 18%) and non-cancerous codes (*n* = 4, 14%). It was not clear from the available information precisely how these various classes of codes were used for case ascertainment (Table 1). Stratification by study type indicated a possible difference in code inclusion between study types (Table 1). Both incidence studies included non-malignant codes (borderline, suspected,

benign, and non-cancerous). Although lists were not received for 2 other incidence studies, they both stated using non-malignant codes within their publication.<sup>21,22</sup> In contrast, only 14/26 (56%) drug safety and epidemiological studies included non-malignant codes (Table 2).

#### Identification and validation of cancers

Of the 84 studies, the majority (*n* = 57) only required ≥1 cancer diagnosis Read code to identify cases (Table 2). Twenty-seven studies specified additional criteria to confirm case status, for example, chemo-radiotherapy (*n* = 13), biological treatment (*n* = 12), and surgical procedures (*n* = 13). The requirement for further evidence was more common for breast cancer studies (76%) compared to colorectal (44%) and prostate cancer studies (50%). Eleven studies mentioned a manual review process but did not report the criteria used to confirm or refute case status. Where present, descriptions of diagnostic algorithms were typically brief, and only one study provided a schematic of the algorithm used to identify and confirm case status.<sup>21</sup> Few studies (4/27) reported on the proportion of cases included once additional confirmatory evidence was applied. Gonzalez-Perez *et al.*<sup>23</sup> reported that 3708/3886 (95.4%) incident breast cancer cases had supporting evidence of diagnosis. Charlton *et al.*<sup>21</sup> identified 1809 potential colorectal cancer cases,

Table 2. Criteria used to identify, validate, and exclude potential cancer cases by cancer and study type

	All studies	Cancer type*			Study type		
		Breast	Colorectal	Prostate	Drug-cancer	Disease-cancer	Incidence
		n (column %, n/N)					
Total number of studies	N = 84	N = 51	N = 54	N = 31	N = 51	N = 28	N = 5
Number of studies requiring ≥1 cancer diagnosis code only	57 (68)	34 (67)	45 (83)	23 (74)	33 (65)	21 (75)	3 (60)
Internal validation or requirement for supportive evidence of diagnosis: number of studies	N = 27	N = 17	N = 9	N = 8	N = 18	N = 7	N = 2
Cancer related surgery	13 (48)	11 (65)	4 (44)	3 (38)	8 (44)	3 (43)	2 (100)
Chemo/radiotherapy	13 (48)	10 (59)	4 (44)	4 (50)	7 (39)	4 (57)	2 (100)
Biological treatment	12 (44)	10 (59)	2 (22)	4 (50)	8 (44)	3 (43)	1 (50)
Treatment unspecified	1 (4)	1 (6)	0 (0)	0 (0)	1 (6)	0 (0)	0 (0)
Consultation with oncologist	8 (30)	7 (41)	2 (22)	1 (13)	6 (31)	1 (14)	1 (50)
Other †	3 (11)	2 (12)	3 (33)	1 (13)	1 (6)	1 (14)	1 (50)
Unspecified‡	11 (41)	4 (24)	5 (56)	4 (50)	8 (44)	3 (43)	0 (0)
Cancer related exclusion criteria: number of studies							
Previous diagnosis of any cancer	43 (51)	31 (61)	25 (46)	19 (61)	29 (57)	12 (43)	2 (40)
Previous diagnosis of cancer of interest	25 (30)	10 (20)	18 (33)	6 (19)	16 (31)	6 (21)	3 (60)
Time related exclusion periods	59 (70)	32 (63)	39 (72)	24 (77)	37 (37)	19 (68)	3 (60)

\*One study could contribute to >1 study type.

†Other includes: specific oncology codes, terminal illness, palliative care, and death within 180 days of diagnosis.

‡A manual review process was conducted; however, criteria used to confirm case status was not described.

of which 1599 patients (88.3%) had additional supporting evidence of diagnosis: colorectal cancer related surgery confirmed 927 cases (51.2%) and non-surgical support such as chemo-radiotherapy or palliative care confirmed 278 cases (15.4%). Of note, Bodmer *et al.*<sup>24</sup> assessed the effect of metformin on colorectal cancer incidence within the CPRD. Similar estimates were obtained regardless of the requirement for confirmatory evidence of diagnosis (OR for ≥50 prescriptions vs never use=1.43; 95% CI, 1.08–1.90 when cases were defined by codes alone, and 1.46; 95% CI, 1.03–2.06 when restricting to those with further supportive evidence of cancer).

Fourteen CPRD studies validated a sample of potential cases using information external to the database, namely by GP questionnaire, or through a request of patient records (Table 3). The proportion of confirmed cases was high [Median Positive Predictive Value (PPV)=0.99; Range, 0.90–1.00], although validity measures were limited to PPV; other measures of validity such as sensitivity and specificity were not assessed. The number of potential cases sampled was low [Median % 4.0; Range, 0.8–11.1]. The median proportion of responses received was high [Median proportion 0.95; Range, 0.87–1.00]. External validation results stratified by cancer type were generally similar.

Two studies examined the concordance of recorded cancer diagnosis between the CPRD and UK Cancer Registry (UKCR).<sup>16,25</sup> Estimates of concordance between the CPRD and UKCR were high [Median PPV 0.9; Range, [0.8–0.9]. Dregan *et al.*<sup>16</sup> reported a

PPV of 0.98 for colorectal cancer; and Boggon *et al.*<sup>25</sup> reported similarly high PPVs for cancer of the breast (503/560=0.90); prostate (600/725=0.83); and colorectum (618/681=0.91).

## DISCUSSION

### Overview

This review has revealed several common shortcomings related to the description of methods used to identify cancer cases in UK primary care database studies. We found that few studies reported the methods used to compile code lists, or made code lists available, limiting the reproducibility of studies. Furthermore, where information was available, we observed substantial variation in codes included. High positive predictive estimates were reported for all three cancer types from studies that used information external to the database to validate cases, but other measures of validity such as sensitivity and specificity were not generally explored.

### Accessibility of code lists

Only 11/84 studies made their code lists available in the publication or specifically mentioned that they could be requested. Code lists may not have been made available for several reasons. For the earlier studies included, there may have been no practical way of publishing a long code list. More recently, most journals have begun accepting web appendix materials without space limits, and other alternatives

Table 3. External validation of potential cases by cancer type

	All studies	Cancer type*		
		Breast	Colorectal	Prostate
	<i>n</i> (column %), median [Range], unless otherwise specified			
Total number of studies	<i>N</i> = 84	<i>N</i> = 51	<i>N</i> = 54	<i>N</i> = 31
Number of studies that validated cases externally by questionnaire or request for patient records (%) <sup>‡</sup>	14 (20)	5 (12)	4 (10)	3 (14)
Number of potential cases sampled for external validation	100 [23–200]	114 [30–114]	85 [23–200]	100 [100–100]
Proportion of cases randomly sampled for external validation from patients initially fulfilling inclusion criteria	4.03 [0.81–11.06]	3.07 [0.81–3.07]	6.40 [3.49–11.06]	7.21 [4.58–9.85]
Proportion of responses received	0.95 [0.87–1.00]	0.95 [0.95–1.00]	0.96 [0.87–1.00]	– <sup>†</sup>
Proportion of cases confirmed	0.99 [0.90–1.00]	1.00 [1.00–1.00]	0.95 [0.90–1.00]	0.98 [0.98–0.98]
Number of studies that validated cases externally by linkage to cancer registry (%)	<i>N</i> = 2	<i>N</i> = 1	<i>N</i> = 2	<i>N</i> = 1
Number of potential CPRD cases sampled for external validation	703 [–]	560 [–]	1228 [681–1775]	725 [–]
Proportion of cases confirmed in cancer registry median [range]	0.90 [0.83–0.94]	0.90 [–]	0.94 [0.91–0.98]	0.83 [–]

\*One study can contribute to >1 cancer type.

<sup>†</sup>Only one study reported the number of responses received—Ronquist *et al.*: 88 responses received from a request of 100 patient records.

<sup>‡</sup>Two studies included in 'All studies' but were not included in specific cancer type columns as they externally validated overall cancer—not distinguishing by cancer type.

have emerged, such as including a web link in the paper to a central code lists repository or registry of studies. Making code lists 'available on request' is problematic since there may be difficulties in contacting the original corresponding author, particularly as time elapses after publication. Some authors simply may not have considered code lists to be important supplementary information, suggesting a need to raise awareness of the need for clear reporting of case definitions. Last, there may be some reluctance among researchers to release code lists due to concerns that they could be used by competing research groups and without due credit.

#### Variation in case definitions and code lists

There was considerable variation in the specific codes used by researchers to identify cancers. The Read code dictionary is updated regularly but we did not find this to be an important driver of variation between code lists: the vast majority of codes used by investigators were available throughout the period during which the included studies took place. It is worth noting that variation in code lists will not necessarily translate to an equivalent variation in selected cases, which will also depend on how commonly specific codes are used: for example, if a majority of cases of breast cancer have a Read code for 'Malignant Neoplasm of Female Breast' (B34.00, which was included in all code lists for breast cancer studies) then these cases will be identified regardless of the rest of the code list. In the other direction, including a code which is never used in practice will have no effect on case ascertainment.

As well as variation in individual codes, we also found variation in types of codes included: all lists included definite malignant diagnosis codes, but some included other code types such as in-situ neoplasm or suspected cancer. Some of the variation in definitions is likely to have arisen from differing study objectives; we noted differences by study type, as may be expected: for example, pharmacoepidemiological studies aiming for high specificity may only include definite malignant neoplasms and exclude borderline codes,<sup>19,26</sup> while incidence studies may use a broad code list to maximise sensitivity, and then attempt to confirm diagnosis in a second stage of review.<sup>21,22</sup> Some studies included benign tumour and non-cancer codes without explanation; whether such codes were included mistakenly or were used specifically to exclude cases is unclear.

The majority of studies required only a cancer diagnosis code as part of their case definition, but around a third of studies required some form of further supportive evidence to confirm case status. Again we found limited details in many study reports on the specific diagnostic algorithms used; one study presented a full schematic illustrating the case definition algorithm,<sup>21</sup> and more routine use of such diagrams might help to improve clarity.

#### External validation of cancer cases

A number of studies validated cases externally by request of patient records or by GP questionnaire, and were generally able to confirm a high proportion of cases. However, not all practices participate in validation or

linkage studies, which may limit the generalisability of validity findings if participating practices differ from non-participating practices in terms of record-keeping practices. It is also unclear whether GP practices asked to validate cases in this way are accessing extra information, or simply referring to the same electronic record used to identify the case, which would inevitably lead to optimistic validity estimates.

#### *Limitations of this review*

This review has several limitations: first, results are limited to cancers of the breast, colorectum, and prostate, and may not apply to other malignancies. Nonetheless, many of the studies included in this review examined multiple cancers, and applied case ascertainment in a global fashion rather than separately for each cancer type. Second, the review was limited to UK primary care database studies; whether the variation observed in this review occurs in non-UK databases is unknown. Last, authors who completed questionnaires and sent code lists may have been a selective group and because of this, their responses may not be generalisable to all researchers. Non-response or an unwillingness to share code lists may have arisen due to concern about methodological criticism, or protectiveness over intellectual property.

#### *Importance and implications*

Our study highlights the variation and lack of transparency in many studies to date on a critical methodological feature of database studies of cancer outcomes, namely the definition and ascertainment of cancer cases. Primary care databases and routine healthcare records are increasingly used in cancer research: we found 84 relevant articles covering just 3 cancer sites; a broader search not restricted by site finds >250 articles including in leading general medical journals and influential specialist journals. Clarity over case ascertainment methods is important for interpreting study findings, reproducing analyses, and understanding the drivers of conflicting or discrepant results.<sup>8,9</sup> Recent work by the Observational Medical Outcomes Partnership has highlighted that design decisions in observational pharmacoepidemiology studies profoundly affect study results,<sup>27</sup> further emphasising the importance of clear and transparent reporting. As well as directly highlighting the need for such transparency and thus influencing future studies, our work can also inform guidelines aimed at improving the quality of reporting for electronic healthcare record research, which are currently in development as part of the RECORD project.<sup>28</sup>

#### *Conclusion*

We comprehensively investigated several aspects of case ascertainment from studies utilising primary care databases for research related to cancer. Methods used to develop case definitions were often unclear, and specific code lists were seldom published or made available. Where provided, we found considerable variation in case definitions and code lists, and the impact of this on case ascertainment is unclear. Future research might clarify the extent to which methodological variations identified in this review impact on findings in applied epidemiological studies, and further explore ways of validating cancer case definitions, including through the use of linked data sources and free-text information.<sup>16,25,29</sup> It is hoped that this study will help to promote clearer reporting of cancer case ascertainment methods, better access to code lists, and a resulting improvement in the transparency and reproducibility of research in this growing field.

#### CONFLICT OF INTEREST

Co-authors (KB, ID, LS) of this manuscript were involved in studies included in this review. However, the classification review process of all received code lists was conducted according to pre-specified criteria for all studies. OK had received unrestricted funding for pharmacoepidemiological research from the Dutch private-public funded Top Institute Pharma. In the context of the IMI Joint Undertaking (IMI JU), the Department of Pharmacoepidemiology, Utrecht University, also received a direct financial contribution from Pfizer.

#### KEY POINTS

- Methods used to create outcome code lists were not transparent in the majority of studies included in this review, and the overall accessibility of study code lists was low.
- We found substantial variation in the way cancer cases were defined, including in the specific diagnosis codes used, and the requirements for further confirmatory evidence. This could potentially impact case ascertainment and study findings.
- Cancer outcomes defined using database-recorded information had high positive predictive value, when validated against external data sources, but few data were available on other measures of validity such as sensitivity and specificity.
- Transparency and reproducibility of research would be improved by clearer reporting of methods used to develop case definitions, and by making code lists available for all published studies.

## ETHICS STATEMENT

The authors state that no ethical approval was needed.

## ACKNOWLEDGEMENTS

The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium, [www.imi-protect.eu](http://www.imi-protect.eu)) which is a public-private partnership coordinated by the European Medicines Agency. The PROTECT project has received support from the Innovative Medicines Initiative Joint Undertaking ([www.imi.europa.eu](http://www.imi.europa.eu)) under grant agreement no. 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution. The views expressed are those of the authors only. KB is funded by an NIHR postdoctoral fellowship (PDF-2011-04-007).

## REFERENCES

- Li Y, Mann RD. The VAMP Research multi-purpose database in the U.K. *Journal of Clinical Epidemiology* 1995; **48**(3): 431–443.
- Clinical Practice Research Datalink Bibliography (internet). [www.cprd.com/bibliography/](http://www.cprd.com/bibliography/) [Accessed 10 April 2013].
- The Health Improvement Network Bibliography (internet). <http://cdmruk.cegedim.com/THINBibliography.pdf> [Accessed 10 April 2013].
- QRESEARCH Bibliography (internet). <http://www.qresearch.org/SitePages/publications.aspx> [Accessed 10 April 2013].
- Chisholm J. The Read clinical classification. *BMJ* 1990; **300**(6732): 1092.
- Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiology and Drug Safety* 2009; **18**(8): 704–707.
- Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. *PLoS ONE* 2009; **4**(9): e7168.
- de Vries F, de Vries C, Cooper C, Leufkens B, van Staa TP. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the General Practice Research Database. *International Journal of Epidemiology* 2006; **35**(5): 1301–1308.
- Dixon WG, Solomon DH. Bisphosphonates and esophageal cancer—a pathway through the confusion. *Nature Reviews. Rheumatology* 2011; **7**(6): 369–372.
- (ONS) OfNS. Cancer Statistics Registrations, England (Series MB1), No. 41, 2010. <http://www.ons.gov.uk/ons/relation/1/cancer-statistics-registrations-england-series-mb1-no-41-2010/index.html> [Accessed 1 March 2013].
- Boston Collaborative Drug Surveillance Program. <http://www.bu.edu/bcdsp/publications-2/> [Accessed 10 April 2013].
- General Practice Notebook - a UK medical reference (internet). <http://www.gpnotebook.co.uk/> [Accessed 1 March 2013].
- Patient.co.uk (internet). [www.patient.co.uk](http://www.patient.co.uk) [Accessed 10 April 2013].
- Fritz A, Percy C, Jack A, Sharmugaratnam K, Sobin L, Parkin DM, Whelan S. International Classification of Diseases for Oncology (Ed 3). Geneva, Switzerland: World Health Organization, 2000.
- World Health O. ICD-10: international statistical classification of diseases and related health problems. World Health Organization: Geneva, 2004.
- Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer Epidemiology* 2012; **36**(5): 425–429.
- Mackenzie IS, Macdonald TM, Thompson A, Morant S, Wei L. Spironolactone and risk of incident breast cancer in women older than 55 years: retrospective, matched cohort study. *BMJ* 2012; **345**: e4447.
- Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to cyclooxygenase-2 inhibitors and risk of cancer: Nested case-control studies. *British Journal of Cancer* 2011; **105**(3): 452–459.
- Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to statins and risk of common cancers: a series of nested case-control studies. *BMC Cancer* 2011; **11**: 409.
- Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to bisphosphonates and risk of cancer: a protocol for nested case-control studies using the QResearch primary care database. *BMJ Open* 2012; **2**(1): e000548.
- Charlton R, Snowball J, Bloomfield K, de Vries C. Colorectal cancer incidence on the General Practice Research Database. *Pharmacoepidemiology and Drug Safety* 2012; **21**(7): 775–783.
- Kaye JA, Derby LE, del Mar Melero-Montes M, Quinn M, Jick H. The incidence of breast cancer in the General Practice Research Database compared with national cancer registration data. *British Journal of Cancer* 2000; **83**(11): 1556–1558.
- Gonzalez-Perez A, Garcia Rodriguez LA. Breast cancer risk among users of antidepressant medications. *Epidemiology* 2005; **16**(1): 101–105.
- Bodmer M, Becker C, Meier C, Jick SS, Meier CR. Use of metformin is not associated with a decreased risk of colorectal cancer: a case-control analysis. *Cancer Epidemiology, Biomarkers & Prevention* 2012; **21**(2): 280–286.
- Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiology and Drug Safety* 2013; **22**(2): 168–175.
- Bhaskaran K, Douglas I, Evans S, van Staa T, Smeeth L. Angiotensin receptor blockers and risk of cancer: cohort study among people receiving antihypertensive drugs in UK General Practice Research Database. *BMJ* 2012; **344**: e2697.
- Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. *Therapeutic Advances in Drug Safety* 2013; **2042098613477445**.
- Benchimol EI, Langan S, Guttman A, Record Steering Committee. Call to RECORD: the need for complete reporting of research using routinely collected health data. *Journal of Clinical Epidemiology* 2013; **66**(7): 703–705.
- Tate AR, Martin AG, Ali A, Cassell JA. Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer. *BMJ Open* 2011; **1**(1): e000025.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

## 9 Appendix B: Supplementary tables for Chapter 3

Table 9.1: Detailed summary of bias assessment by study

Author & Year	Did the study suffer potentially from the following biases?				
	Immortal Time Bias	Protopathic Bias	Prevalent user bias	Healthy user bias	Time-window bias
Chang 2011 <sup>68</sup>	<b>Not examined:</b> case-control study used	<b>No:</b> lag time of 1-year was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> adjusted for number of physician visits and number of hospitalisations	<b>No:</b> Controls matched on case date of diagnosis
Farwell 2011 <sup>132</sup>	<b>No:</b> follow-up period for both treatment groups started after 2 years of statin exposure.	<b>No:</b> lag time of 2-years was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Active comparative group consisted of antihypertensive medication users used.	<b>Not examined:</b> Cohort design used
Tan 2011 <sup>133</sup>	<b>No:</b> immortal time excluded	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> all patients underwent a PSA test or digital rectal examination	<b>Not examined:</b> Cohort design used
Vinogradova 2011 <sup>130</sup>	<b>Not examined:</b> case-control study used	<b>No:</b> lag time of 1-year was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>No:</b> risk-set sampling used to select controls
Hippisley-Cox 2010 <sup>131</sup>	<b>Uncertain:</b> unequal start dates between treatment groups	<b>Yes:</b> no minimum period of exposure was implemented.	<b>No:</b> new user design used for statin group	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>Not examined:</b> Cohort design used
Murtola 2010 <sup>134</sup>	<b>No:</b> Cox-proportional hazards with time dependent exposure was used.	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> PSA screening arm of Finnish trial utilised as cohort of statin and non-statin users.	<b>Not examined:</b> Cohort design used
Robertson 2010 <sup>135</sup>	<b>Not examined:</b> case-control study used.	<b>No:</b> lag time of 2-years was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>No:</b> risk-set sampling used to select controls
Wooditschka 2010 <sup>136</sup>	<b>Not examined:</b> case-control study used.	<b>No:</b> lag time of 101 days was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>No:</b> Controls matched to cases on duration of follow-up
Flick 2009 <sup>137</sup>	<b>No:</b> Cox-proportional hazards with time dependent exposure was used.	<b>No:</b> a 101 day minimum period of statin use was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> adjusted for history of sigmoidoscopy.	<b>Not examined:</b> Cohort design used
Haukka 2009 <sup>67</sup>	<b>Uncertain:</b> start of follow-up not defined for non-users	<b>No:</b> lag time of 1-year was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>Not examined:</b> Cohort design used
Singh 2009 <sup>64</sup>	<b>Yes:</b> ≥2 statins exposure definition. However, follow-up started at the first statin, introducing immortal time.	<b>Yes:</b> no minimum period of exposure was implemented.	<b>No:</b> sensitivity analysis including only <i>new statin</i> users	<b>No:</b> adjusted for lower GI endoscopy	<b>Not examined:</b> Cohort design used
Boudreau 2008 <sup>139</sup>	<b>No:</b> Cox-proportional hazards with time dependent exposure was used	<b>No:</b> lag time of 1-year was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Conducted secondary analyses of men who had a PSA test within 5 years of study start.	<b>Not examined:</b> Cohort design used
Boudreau 2008 <sup>140</sup>	<b>Not examined:</b> case-control study used.	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>No:</b> risk-set sampling used to select controls

[Table 9.1 continued over]

[Table 9.1 continued]					
Did the study suffer potentially from the following biases?					
Author & Year	Immortal Time Bias	Protopathic Bias	Prevalent user bias	Healthy user bias	Time-window bias
Farwell 2008 <sup>141</sup>	<b>No:</b> follow-up period for both treatment groups started after 2 years of statin exposure.	<b>No:</b> lag time of 2-years was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Active comparative group consisted of antihypertensive medication users used.	<b>Not examined:</b> Cohort design used
Friedman 2008 <sup>142</sup>	<b>No:</b> Cox-proportional hazards with time dependent exposure was used.	<b>No:</b> lag time of 2-years was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made. <b>No (prostate cancer):</b> PSA adjusted	<b>Not examined:</b> Cohort design used
Hachem 2008 <sup>138</sup>	<b>Not examined:</b> case-control study used.	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> adjusted for colorectal evaluation by imaging, endoscopy, and faecal occult blood testing	<b>No:</b> Controls matched to cases on duration of follow-up
Smeeth 2008 <sup>56</sup>	<b>No:</b> immortal time excluded	<b>No:</b> lag time of 2-years was implemented.	<b>No:</b> new user design used for statin group	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>Not examined:</b> Cohort design used
Yang 2008 <sup>127</sup>	<b>Not examined:</b> case-control study used.	<b>No:</b> lag time of 5-years was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> adjustment for history of colonoscopy or sigmoidoscopy	<b>No:</b> Controls matched to cases on duration of follow-up
Boudreau 2007 <sup>143</sup>	<b>No:</b> Cox-proportional hazards with time dependent exposure was used.	<b>No:</b> lag time of 1-year was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> adjusted for breast cancer screening	<b>Not examined:</b> Cohort design used
Flick 2007 <sup>144</sup>	<b>No:</b> Cox-proportional hazards with time dependent exposure was used.	<b>No:</b> a 101 day minimum period of statin use was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Conducted secondary analyses of men who had a PSA test within 5 years of study start.	<b>Not examined:</b> Cohort design used
Murtola 2007 <sup>145</sup>	<b>Not examined:</b> case-control study used.	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>No:</b> risk-set sampling used to select controls
Vinogradova 2007 <sup>90</sup>	<b>Not examined:</b> case-control study used.	<b>No:</b> lag time of 1-year was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>No:</b> risk-set sampling used to select controls
Khurana 2007	<b>Not examined:</b> case-control study used.	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>Yes:</b> time-independent sampling of controls
Setoguchi 2006 <sup>146</sup>	<b>No:</b> immortal time excluded	<b>No:</b> lag time of 6-months was implemented.	<b>No:</b> New user design implemented for both statin users and glaucoma medication users	<b>No:</b> Active comparative group consisted of glaucoma medication groups	<b>Not examined:</b> Cohort design used
Friis 2004 <sup>147</sup>	<b>Uncertain:</b> follow-up not defined	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Comparative group consisted of other lipid-lowering drug users.	<b>Not examined:</b> Cohort design used
Graaf 2004 <sup>148</sup>	<b>Not examined:</b> case-control study used.	<b>No:</b> lag time of 6-months implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Comparative group consisted of antihypertensive drug users.	<b>No:</b> risk-set sampling used to select controls
Kaye 2004 <sup>128</sup>	<b>Not examined:</b> case-control study used.	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Comparative group consisted of other lipid-lowering drug users.	<b>No:</b> risk-set sampling used to select controls
Beck 2003 <sup>149</sup>	<b>No:</b> start of follow-up at first prescription – immortal time excluded	<b>Yes:</b> no minimum period of exposure was implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>Yes:</b> comparison group of non-users used and no adjustment for health service utilisation was made.	<b>Not examined:</b> Cohort design used

[Table 9.1 continued over]



[Table 9.1 continued]					
Did the study suffer potentially from the following biases?					
Author & Year	Immortal Time Bias	Protopathic Bias	Prevalent user bias	Healthy user bias	Time-window bias
Kaye 2002 <sup>129</sup>	<b>Not examined:</b> case-control study used.	<b>Yes:</b> No minimum period of exposure	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Comparative group consisted of other lipid-lowering drug users.	<b>No:</b> risk-set sampling used to select controls
Blais 2000 <sup>150</sup>	<b>Not examined:</b> case-control study used.	<b>No:</b> lag time of 1-year implemented.	<b>Yes:</b> both new and prevalent statin users included	<b>No:</b> Comparison group: bile acid-binding resins	<b>No:</b> risk-set sampling used to select controls
GI: Gastro-intestinal; PSA: Prostate-specific Antigen					

**Table 9.2: Updated review studies: Statin use associated with cancer risk - study details**

Author & Year	Cancer(s) examined	Study Population	Data source for outcomes	Invasive or non-invasive cancer	Data source for statin use	Time period	Study design	Comparison group	Statin exposure definition	No of cases (cases exposed)	Total cancer or site-specific cancer: point estimate (95% CI)
Nordstrom 2015	Prostate	Stockholm County undergoing a PSA test	National Prostate Cancer registry	Invasive	Swedish prescribed drug register	2007-2012	Cohort	Non-Statin users	≥1 statin prescription in 2 years prior to cohort entry	8430 (-)	1.16 (1.04, 1.29)
Bjorkhem-Bergman 2014	Colon	Swedish population register	Swedish cancer registry	Invasive	Swedish prescribed drug register	2006-2010	Case-control	Non-Statin users	Three registrations of dispensed statins in a 1 year period	21143 (4218)	1.04 (1.00, 1.08)
Chan 2014	Breast	Taiwan National Health Insurance program	Insurance database	Invasive	Prescription database	1996-2010	Case-control	Non-Statin users	≥1 statin prescription at any time during the study period	565 (130)	1.13 (0.84, 1.51)
Jesperperson 2014	Prostate	Danish Civil Registration System	Cancer registry	Invasive	Danish prescription registry	1996-2010	Case-control	Non-Statin users	≥1 statin prescription 0-12 months prior to date of diagnosis	42480 (7125)	0.94 (0.91, 0.97)
Lustman 2014	Prostate	Clait health services registered patients; Israel	Disease register	Invasive	Prescription database	2000-2009	Cohort	Non-Statin users	≥1 statin prescription between 2001-2009	306 (-)	0.68 (0.60, 0.79)

[Table 9.2 continued over]

[Table 9.2 continued]

Author & Year	Cancer(s) examined	Study Population	Data source for outcomes	Invasive or non-invasive cancer	Data source for statin use	Time period	Study design	Comparison group	Statin exposure definition	No of cases (cases exposed)	Total cancer or site-specific cancer: point estimate (95% CI)
Clancy 2013	Colorectal	Northern Italy -Emilia-Romagna Region	Hospital discharge data	Invasive	Outpatient pharmacy data	2003-2009	Cohort	Glaucoma medication users	New user: 3 filled prescriptions within 180days of the first prescription	1870 (215963)	0.84 (0.76, 0.92)
Lakha 2012	Colorectal	Tayside, Scotland	Hospital records	Invasive	Linked pharmacy data	1996-2006	Case-control	Non-Statin users	≥1 statin prescription at 2 or 6 months prior to date of diagnosis	309 (25)	0.49 (0.22, 1.08)
Lutski 2012	Breast Colorectal Prostate Lung	Maccabi healthcare services, Israel	Israel cancer registry	Invasive	Prescription database	1998-2007	Cohort	Non-Statin users	≥1 year continuous statin exposure	8662 (-)	0.90 (0.84–0.96)
Cheng 2012	Lung	Taiwan National Health Insurance program	Insurance database	Invasive	Prescription database	2005-2008	Case-control	Non-Statin users	≥1 statin prescription at any time during the study period	297 (61)	0.82 (0.58, 1.15)
Cheng 2011	Colorectal	Taiwan National Health Insurance program	Insurance database	Invasive	Prescription database	1996-2008	Case-control	Non-Statin users	≥1 statin prescription at any time during the study period	1156 (242)	1.09 (0.91, 1.30)

**Table 9.3: Summary of Findings and Biases - Risk of breast, prostate, and colorectal and lung cancer associated with Statin Use**

Author & Year	Cancer type	No of cases (cases exposed)	Immortal	Protopathic	Prevalent user	Healthy user	Time-window	Confounders	Analysis method	Main result (95% CI)
Nordstrom 2015	Prostate	8430 (-)	Yes	Yes	Yes	No	-	age, psa (log transformed), psa quotient, Charlston morbidity index, education, aspirin use, antidiabetic medications	Logistic regression	1.24 (1.10, 1.42)
Bjorkhem-Bergman 2014	Colon	21143 (4218)	-	Yes	Yes	Yes	No	age, sex, diabetes, education, cortisone, acetylsalicylic acid, NSAID, chemotherapy, chron's disease, ulcerous colitis	Cox regression	1.04 (1.00, 1.08)
Chan 2014	Breast	565 (130)	-	No	Yes	No	No	age, sex, index date, benign mammary dysphasia, mammography, NSAIDs	Conditional logistic regression	1.13 (0.84, 1.51)
Jesperon 2014	Prostate	42480 (7125)	-	Yes	Yes	Yes	No	age, level of comorbidity, aspirin use, NSIAD use, education	Conditional logistic regression	0.94 (0.91, 0.97)
Lustman 2014	Prostate	306 (-)	No	Yes	Yes	No	-	age, diabetes, BMI, CVD, smoking status	Time dependent Cox regression	0.68 (0.60, 0.79)
Clancy 2013	Colorectal	1870 (215963)	No	No	No	No	-	age, sex, colonoscopy, bowel disease, NSAID, estrogen, obesity, no. of co-medications, no. of chronic conditions, no. of hospitalisations, Charleston weighted index	Cox regression	0.84 (0.76, 0.92)
Lakha 2012	Colorectal	309 (25)	-	No	Yes	Yes	U	age, sex, region, family history of cancer, history of cancer, bowel disease, BMI, smoking, physical activity, NSAID	Conditional logistic regression	0.49 (0.22, 1.08)
Lutski 2012	Breast Colorectal Prostate	8662 (-)	No	No	No	No	-	age, sex, marital status, area of residence, nationality, socioeconomic level, years of stay in Israel, obesity, diabetes mellitus, hypertension, cardiovascular disease, efficacy, hospitalizations and visits to physicians a year before first statin dispensation, as well as for asthma, and chronic obstructive pulmonary disease (for lung and bronchus cancers).	Cox regression	breast: 1.03 (0.87, 1.22) colorectal: 0.93 (0.78, 1.11) prostate: 0.95 (0.81, 1.13)
Cheng 2012	Lung		-	Yes	Yes	Yes	No	Adjusted for matching variable, tuberculosis, diabetes, use of NSAID, HRT use, and use of other lipid-lowering drugs.	Conditional logistic regression	0.82 (0.58, 1.15)
Cheng 2011	Colorectal	1156 (242)	-	Yes	Yes	No	No	age, sex, index date, diabetes, number of hospitalisations, cholecystectomy, liver disease, colorectal polyps, infammatory bowel disease, colonoscopy, fecal occult blood testing, NSAIDs, other lipid lowering drugs	Conditional logistic regression	1.09 (0.91, 1.30)

## **10 Appendix C: Supplementary tables for Chapter 4**

### **Appendix C.1: ISAC and LSHTM Ethics approval details**

ISAC approval (application number 12\_068R) for this project was obtained in 15 August 2013. London School of Hygiene & Tropical Medicine (LSHTM) ethics approval for this study was obtained on 29 May 2012 (Application Number 6202).

**Table 10.1: Breast cancer code list**

Readcode	Description	Classification	ICD-10 code
B342.00	MALIGNANT NEOPLASM OF UPPER-INNER QUADRANT OF FEMALE BREAST	Malignant Neoplasm	C50
B34z.00	MALIGNANT NEOPLASM OF FEMALE BREAST NOS	Malignant Neoplasm	C50
B34yz00	MALIGNANT NEOPLASM OF OTHER SITE OF FEMALE BREAST NOS	Malignant Neoplasm	C50
B340z00	MALIGNANT NEOPLASM OF NIPPLE OR AREOLA OF FEMALE BREAST NOS	Malignant Neoplasm	C50
B346.00	MALIGNANT NEOPLASM OF AXILLARY TAIL OF FEMALE BREAST	Malignant Neoplasm	C50
B340000	MALIGNANT NEOPLASM OF NIPPLE OF FEMALE BREAST	Malignant Neoplasm	C50
B340.00	MALIGNANT NEOPLASM OF NIPPLE AND AREOLA OF FEMALE BREAST	Malignant Neoplasm	C50
B34y000	MALIGNANT NEOPLASM OF ECTOPIC SITE OF FEMALE BREAST	Malignant Neoplasm	C50
Byu6.00	[X]MALIGNANT NEOPLASM OF BREAST	Malignant Neoplasm	C50
B344.00	MALIGNANT NEOPLASM OF UPPER-OUTER QUADRANT OF FEMALE BREAST	Malignant Neoplasm	C50
B340100	MALIGNANT NEOPLASM OF AREOLA OF FEMALE BREAST	Malignant Neoplasm	C50
B347.00	MALIGNANT NEOPLASM, OVERLAPPING LESION OF BREAST	Malignant Neoplasm	C50
B34y.00	MALIGNANT NEOPLASM OF OTHER SITE OF FEMALE BREAST	Malignant Neoplasm	C50
B34..00	MALIGNANT NEOPLASM OF FEMALE BREAST	Malignant Neoplasm	C50
B34..11	CA FEMALE BREAST	Malignant Neoplasm	C50
B341.00	MALIGNANT NEOPLASM OF CENTRAL PART OF FEMALE BREAST	Malignant Neoplasm	C50
B343.00	MALIGNANT NEOPLASM OF LOWER-INNER QUADRANT OF FEMALE BREAST	Malignant Neoplasm	C50
B345.00	MALIGNANT NEOPLASM OF LOWER-OUTER QUADRANT OF FEMALE BREAST	Malignant Neoplasm	C50
BB9K000	[M]PAGET'S DISEASE AND INTRADUCTAL CARCINOMA OF BREAST	Malignant Morphology	C50
BB9G.00	[M]INFILTRATING DUCTULAR CARCINOMA	Malignant Morphology	C50
BB91.00	[M]INFILTRATING DUCT CARCINOMA	Malignant Morphology	C50
BB91000	[M]INTRADUCTAL PAPILLARY ADENOCARCINOMA WITH INVASION	Malignant Morphology	C50

Table 10.1 continued over

Table 10.1 continued

Read/OXMIS code	Description	Classification	
BB9H.00	[M]INFLAMMATORY CARCINOMA	Malignant Morphology	C50
BB94.11	[M]SECRETORY BREAST CARCINOMA	Malignant Morphology	C50
BB9J.11	[M]PAGET'S DISEASE, BREAST	Malignant Morphology	C50
BB9F.00	[M]LOBULAR CARCINOMA NOS	Malignant Morphology	C50
BB94.00	[M]JUVENILE BREAST CARCINOMA	Malignant Morphology	C50
BB93.00	[M]COMEDOCARCINOMA NOS	Malignant Morphology	C50
BB9D.00	[M]MEDULLARY CARCINOMA WITH LYMPHOID STROMA	Malignant Morphology	C50
BB91100	[M]INFILTRATING DUCT AND LOBULAR CARCINOMA	Malignant Morphology	C50
BB9K.00	[M]PAGET'S DISEASE AND INFILTRATING BREAST DUCT CARCINOMA	Malignant Morphology	C50
BB9J.00	[M]PAGET'S DISEASE, MAMMARY	Malignant Morphology	C50
BBM9.00	[M]CYSTOSARCOMA PHYLLODES, MALIGNANT	Malignant Morphology	C50
B83..00	CARCINOMA IN SITU OF BREAST AND GENITOURINARY SYSTEM	Malignant in situ tumour	D09
B830.00	CARCINOMA IN SITU OF BREAST	Malignant in situ tumour	D05
B830100	INTRADUCTAL CARCINOMA IN SITU OF BREAST	Malignant in situ tumour	D05
ByuFG00	[X]OTHER CARCINOMA IN SITU OF BREAST	Malignant in situ tumour	D05
B830000	LOBULAR CARCINOMA IN SITU OF BREAST	Malignant in situ tumour	D05
4KJ1.00	PROGESTERONE RECEPTOR POSITIVE TUMOUR	Borderline	D48
4KJ0.00	OESTROGEN RECEPTOR POSITIVE TUMOUR	Borderline	D48
4KJ2.00	OESTROGEN RECEPTOR NEGATIVE TUMOUR	Borderline	D48
BA03.00	NEOPLASM OF UNSPECIFIED NATURE OF BREAST	Borderline	D48
4KJ3.00	PROGESTERONE RECEPTOR NEGATIVE TUMOUR	Borderline	D48
B933.00	NEOPLASM OF UNCERTAIN BEHAVIOUR OF BREAST	Borderline	D48
B58y000	SECONDARY MALIGNANT NEOPLASM OF BREAST	Secondary or metastatic	-

Table 10.1 continued over

Table 10.1 continued

Read/OXMIS code	Description	Classification	
ZV10300	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF BREAST	History of cancer	-
ZV13A00	[V]PERSONAL HISTORY OF NON-NEOPLASTIC BREAST DISEASE	History of cancer	-
1J0I.00	SUSPECTED BREAST CANCER	Suspected	-
6862100	BREAST NEOPLASM SCREEN ABNORM	Suspected	-
8Hn2.00	FAST TRACK REFERRAL FOR SUSPECTED BREAST CANCER	Suspected	-
9Np2.00	SEEN IN FAST TRACK SUSPECTED BREAST CANCER CLINIC	Suspected	-



**Table 10.2: Colorectal cancer code list**

Readcode	Description	Classification	ICD-10 Code
B132.00	MALIGNANT NEOPLASM OF DESCENDING COLON	Malignant Neoplasm	C18
B136.00	MALIGNANT NEOPLASM OF ASCENDING COLON	Malignant Neoplasm	C18
B13z.11	COLONIC CANCER	Malignant Neoplasm	C18
B134.00	MALIGNANT NEOPLASM OF CAECUM	Malignant Neoplasm	C18
B139.00	HEREDITARY NONPOLYPOSIS COLON CANCER	Malignant Neoplasm	C18
B130.00	MALIGNANT NEOPLASM OF HEPATIC FLEXURE OF COLON	Malignant Neoplasm	C18
B135.00	MALIGNANT NEOPLASM OF APPENDIX	Malignant Neoplasm	C18
9Ow1.00	BOWEL CANCER DETECTED BY NATIONAL SCREENING PROGRAMME	Malignant Neoplasm	C18
B1z0.11	CANCER OF BOWEL	Malignant Neoplasm	C18
B13z.00	MALIGNANT NEOPLASM OF COLON NOS	Malignant Neoplasm	C18
B131.00	MALIGNANT NEOPLASM OF TRANSVERSE COLON	Malignant Neoplasm	C18
B13y.00	MALIGNANT NEOPLASM OF OTHER SPECIFIED SITES OF COLON	Malignant Neoplasm	C18
B13..00	MALIGNANT NEOPLASM OF COLON	Malignant Neoplasm	C18
B133.00	MALIGNANT NEOPLASM OF SIGMOID COLON	Malignant Neoplasm	C18
B138.00	MALIGNANT NEOPLASM, OVERLAPPING LESION OF COLON	Malignant Neoplasm	C18
B137.00	MALIGNANT NEOPLASM OF SPLENIC FLEXURE OF COLON	Malignant Neoplasm	C18
B134.11	CARCINOMA OF CAECUM	Malignant Neoplasm	C18
4M10.00	DUKES STAGE A	Malignant Neoplasm	C18
4M11.00	DUKES STAGE B	Malignant Neoplasm	C18
4M12.00	DUKES STAGE C1	Malignant Neoplasm	C18
4M13.00	DUKES STAGE C2	Malignant Neoplasm	C18
4M14.00	DUKES STAGE D	Malignant Neoplasm	C18

Table 10.2 continued over

Table 10.2 continued

Readcode	Description	Classification	ICD-10 Code
4M1..00	DUKES STAGING SYSTEM	Malignant Neoplasm	C18
B140.00	MALIGNANT NEOPLASM OF RECTOSIGMOID JUNCTION	Malignant Neoplasm	C19
B141.11	CARCINOMA OF RECTUM	Malignant Neoplasm	C20
B141.12	RECTAL CARCINOMA	Malignant Neoplasm	C20
B141.00	MALIGNANT NEOPLASM OF RECTUM	Malignant Neoplasm	C20
B14z.00	MALIGNANT NEOPLASM RECTUM,RECTOSIGMOID JUNCTION AND ANUS NOS	Malignant Neoplasm	C21
B142000	MALIGNANT NEOPLASM OF CLOACOGENIC ZONE	Malignant Neoplasm	C21
B142.00	MALIGNANT NEOPLASM OF ANAL CANAL	Malignant Neoplasm	C21
B14..00	MALIGNANT NEOPLASM OF RECTUM, RECTOSIGMOID JUNCTION AND ANUS	Malignant Neoplasm	C21
B142.11	ANAL CARCINOMA	Malignant Neoplasm	C21
B14y.00	MALIG NEOP OTHER SITE RECTUM, RECTOSIGMOID JUNCTION AND ANUS	Malignant Neoplasm	C21
B143.00	MALIGNANT NEOPLASM OF ANUS UNSPECIFIED	Malignant Neoplasm	C21
B18y200	MALIGNANT NEOPLASM OF MESORECTUM	Malignant Neoplasm	C48
BB5N100	[M]ADENOCARCINOMA IN ADENOMATOUS POLPOSIS COLI	Malignant Morphology	C18
BB5R600	[M]MUCOCARCINOID TUMOUR, MALIGNANT	Malignant Morphology	C18
B803700	CARCINOMA IN SITU OF SPLENIC FLEXURE OF COLON	Malignant in situ tumour	D01
B803200	CARCINOMA IN SITU OF DESCENDING COLON	Malignant in situ tumour	D01
B803600	CARCINOMA IN SITU OF ASCENDING COLON	Malignant in situ tumour	D01
B803z00	CARCINOMA IN SITU OF COLON NOS	Malignant in situ tumour	D01
B803000	CARCINOMA IN SITU OF HEPATIC FLEXURE OF COLON	Malignant in situ tumour	D01
B804.00	CARCINOMA IN SITU OF RECTUM AND RECTOSIGMOID JUNCTION	Malignant in situ tumour	D01
B804100	CARCINOMA IN SITU OF RECTUM	Malignant in situ tumour	D01
B803100	CARCINOMA IN SITU OF TRANSVERSE COLON	Malignant in situ tumour	D01

Table 10.2 continued over

Table 10.2 continued

Readcode	Description	Classification	ICD-10 Code
B804z00	CARCINOMA IN SITU OF RECTUM OR RECTOSIGMOID JUNCTION NOS	Malignant in situ tumour	D01
B803.00	CARCINOMA IN SITU OF COLON	Malignant in situ tumour	D01
B804000	CARCINOMA IN SITU OF RECTOSIGMOID JUNCTION	Malignant in situ tumour	D01
B803300	CARCINOMA IN SITU OF SIGMOID COLON	Malignant in situ tumour	D01
B902500	NEOPLASM OF UNCERTAIN BEHAVIOUR OF RECTUM	Borderline	D37
BB5N.00	[M]ADENOMATOUS AND ADENOCARCINOMATOUS POLYPS OF COLON	Borderline	D37
BB5Nz00	[M]ADENOMATOUS OR ADENOCARCINOMATOUS POLYPS OF THE COLON NOS	Borderline	D37
B902400	NEOPLASM OF UNCERTAIN BEHAVIOUR OF COLON	Borderline	D37
B902.00	NEOP OF UNCERTAIN BEHAVIOUR STOMACH, INTESTINES AND RECTUM	Borderline	D37
B575.00	SECONDARY MALIGNANT NEOPLASM OF LARGE INTESTINE AND RECTUM	Secondary or metastatic	C78
B575100	SECONDARY MALIGNANT NEOPLASM OF RECTUM	Secondary or metastatic	C78
B575z00	SECONDARY MALIG NEOP OF LARGE INTESTINE OR RECTUM NOS	Secondary or metastatic	C78
B575000	SECONDARY MALIGNANT NEOPLASM OF COLON	Secondary or metastatic	C78
ZV10017	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF RECTUM	History of cancer	C20
8Hn4.00	FAST TRACK REFERRAL FOR SUSPECTED COLORECTAL CANCER	Suspected	-
8CAo.00	PATIENT GIVEN ADVICE ABOUT BOWEL CANCER	Suspected	-
9Np7.00	SEEN IN FAST TRACK SUSPECTED COLORECTAL CANCER CLINIC	Suspected	-

**Table 10.3: Prostate cancer code list**

Readcode	Description	Classification	ICD-10 Code
4M01.00	GLEASON PROSTATE GRADE 5-7 (MEDIUM)	Malignant Neoplasm	C61
4M02.00	GLEASON PROSTATE GRADE 8-10 (HIGH)	Malignant Neoplasm	C61
4M00.00	GLEASON PROSTATE GRADE 2-4 (LOW)	Malignant Neoplasm	C61
4M0..00	GLEASON GRADING OF PROSTATE CANCER	Malignant Neoplasm	C61
B46..00	MALIGNANT NEOPLASM OF PROSTATE	Malignant Neoplasm	C61
B915.00	NEOPLASM OF UNCERTAIN BEHAVIOUR OF PROSTATE	Borderline	D40
B834.00	CARCINOMA IN SITU OF PROSTATE	Malignant in situ tumour	D07
B58y500	SECONDARY MALIGNANT NEOPLASM OF PROSTATE	Secondary or metastatic	-
ZV10415	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF PROSTATE	Secondary or metastatic	-
1427000	H/O: PROSTATE CANCER	History of cancer	-
1J08.00	SUSPECTED PROSTATE CANCER	Suspected	-

**Table 10.4: Lung Cancer Code list**

Readcode	Description	Classification	ICD-10 Code
B22..00	MALIGNANT NEOPLASM OF TRACHEA, BRONCHUS AND LUNG	Malignant Neoplasm	C34
B220.00	MALIGNANT NEOPLASM OF TRACHEA	Malignant Neoplasm	C33
B220100	MALIGNANT NEOPLASM OF MUCOSA OF TRACHEA	Malignant Neoplasm	C33
B220z00	MALIGNANT NEOPLASM OF TRACHEA NOS	Malignant Neoplasm	C33
B221.00	MALIGNANT NEOPLASM OF MAIN BRONCHUS	Malignant Neoplasm	C34
B221000	MALIGNANT NEOPLASM OF CARINA OF BRONCHUS	Malignant Neoplasm	C34
B221100	MALIGNANT NEOPLASM OF HILUS OF LUNG	Malignant Neoplasm	C34
B221z00	MALIGNANT NEOPLASM OF MAIN BRONCHUS NOS	Malignant Neoplasm	C34
B222.00	MALIGNANT NEOPLASM OF UPPER LOBE, BRONCHUS OR LUNG	Malignant Neoplasm	C34
B222.11	PANCOAST'S SYNDROME	Malignant Neoplasm	C34
B222000	MALIGNANT NEOPLASM OF UPPER LOBE BRONCHUS	Malignant Neoplasm	C34
B222100	MALIGNANT NEOPLASM OF UPPER LOBE OF LUNG	Malignant Neoplasm	C34
B222z00	MALIGNANT NEOPLASM OF UPPER LOBE, BRONCHUS OR LUNG NOS	Malignant Neoplasm	C34
B223.00	MALIGNANT NEOPLASM OF MIDDLE LOBE, BRONCHUS OR LUNG	Malignant Neoplasm	C34
B223000	MALIGNANT NEOPLASM OF MIDDLE LOBE BRONCHUS	Malignant Neoplasm	C34
B223100	MALIGNANT NEOPLASM OF MIDDLE LOBE OF LUNG	Malignant Neoplasm	C34
B223z00	MALIGNANT NEOPLASM OF MIDDLE LOBE, BRONCHUS OR LUNG NOS	Malignant Neoplasm	C34
B224.00	MALIGNANT NEOPLASM OF LOWER LOBE, BRONCHUS OR LUNG	Malignant Neoplasm	C34
B224000	MALIGNANT NEOPLASM OF LOWER LOBE BRONCHUS	Malignant Neoplasm	C34
B224100	MALIGNANT NEOPLASM OF LOWER LOBE OF LUNG	Malignant Neoplasm	C34
B224z00	MALIGNANT NEOPLASM OF LOWER LOBE, BRONCHUS OR LUNG NOS	Malignant Neoplasm	C34
B225.00	MALIGNANT NEOPLASM OF OVERLAPPING LESION OF BRONCHUS & LUNG	Malignant Neoplasm	C34
B22y.00	MALIGNANT NEOPLASM OF OTHER SITES OF BRONCHUS OR LUNG	Malignant Neoplasm	C34
B22z.00	MALIGNANT NEOPLASM OF BRONCHUS OR LUNG NOS	Malignant Neoplasm	C34
Table 10.4 continued over			

Table 10.4 continued			
Readcode	Description	Classification	ICD-10 Code
B22z.11	LUNG CANCER	Malignant Neoplasm	C34
B23..00	MALIGNANT NEOPLASM OF PLEURA	Malignant Neoplasm	C38
B230.00	MALIGNANT NEOPLASM OF PARIETAL PLEURA	Malignant Neoplasm	C38
B23y.00	MALIGNANT NEOPLASM OF OTHER SPECIFIED PLEURA	Malignant Neoplasm	C38
B23z.00	MALIGNANT NEOPLASM OF PLEURA NOS	Malignant Neoplasm	C38
Byu2000	[X]MALIGNANT NEOPLASM OF BRONCHUS OR LUNG, UNSPECIFIED	Malignant Neoplasm	C34
BB1K.00	[M]OAT CELL CARCINOMA	Malignant Morphology	C34
BB1L.00	[M]SMALL CELL CARCINOMA, FUSIFORM CELL TYPE	Malignant Morphology	C34
BB1M.00	[M]SMALL CELL CARCINOMA, INTERMEDIATE CELL	Malignant Morphology	C34
BB1N.00	[M]SMALL CELL-LARGE CELL CARCINOMA	Malignant Morphology	C34
BB5J.12	[M]CYLINDROID BRONCHIAL ADENOMA	Malignant Morphology	C34
BB5R111	[M]CARCINOID BRONCHIAL ADENOMA	Malignant Morphology	C34
BB5S200	[M]BRONCHIOLO-ALVEOLAR ADENOCARCINOMA	Malignant Morphology	C34
BB5S211	[M]ALVEOLAR CELL CARCINOMA	Malignant Morphology	C34
BB5S212	[M]BRONCHIOLAR CARCINOMA	Malignant Morphology	C34
BB5S400	[M]ALVEOLAR ADENOCARCINOMA	Malignant Morphology	C34
B811.00	CARCINOMA IN SITU OF TRACHEA	Malignant in situ tumour	D02
B812.00	CARCINOMA IN SITU OF BRONCHUS AND LUNG	Malignant in situ tumour	D02
B812000	CARCINOMA IN SITU OF CARINA OF BRONCHUS	Malignant in situ tumour	D02
B812100	CARCINOMA IN SITU OF MAIN BRONCHUS	Malignant in situ tumour	D02
B812200	CARCINOMA IN SITU OF UPPER LOBE BRONCHUS AND LUNG	Malignant in situ tumour	D02
B812300	CARCINOMA IN SITU OF MIDDLE LOBE BRONCHUS AND LUNG	Malignant in situ tumour	D02
B812400	CARCINOMA IN SITU OF LOWER LOBE BRONCHUS AND LUNG	Malignant in situ tumour	D02
B812z00	CARCINOMA IN SITU OF BRONCHUS OR LUNG NOS	Malignant in situ tumour	D02
B81y000	CARCINOMA IN SITU OF PLEURA	Malignant in situ tumour	D02
Table 10.4 continued over			

Table 10.4 continued			
Readcode	Description	Classification	ICD-10 Code
B907000	NEOPLASM OF UNCERTAIN BEHAVIOUR OF TRACHEA	Borderline	D38
B907100	NEOPLASM OF UNCERTAIN BEHAVIOUR OF BRONCHUS	Borderline	D38
B907200	NEOPLASM OF UNCERTAIN BEHAVIOUR OF LUNG	Borderline	D38
B907z00	NEOP OF UNCERTAIN BEHAVIOUR OF TRACHEA, BRONCHUS OR LUNG NOS	Borderline	D38
B908000	NEOPLASM OF UNCERTAIN BEHAVIOUR OF PLEURA	Borderline	D38
4D56.00	PLEURAL FLUID: MALIGNANT CELLS	Secondary or metastatic	C79
B570.00	SECONDARY MALIGNANT NEOPLASM OF LUNG	Secondary or metastatic	C79
B572.00	SECONDARY MALIGNANT NEOPLASM OF PLEURA	Secondary or metastatic	C79
H51y700	MALIGNANT PLEURAL EFFUSION	Secondary or metastatic	C79
ZV10100	[V]PERSONAL HISTORY OF MALIG NEOP OF TRACHEA/BRONCHUS/LUNG	History of cancer	C34
ZV10111	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF BRONCHUS	History of cancer	C34
ZV10112	[V]PERSONAL HISTORY OF MALIGNANT NEOPLASM OF LUNG	History of cancer	C34
1J00.00	SUSPECTED LUNG CANCER	Suspected malignancy	-
8Hn7.00	FAST TRACK REFERRAL FOR SUSPECTED LUNG CANCER	Suspected malignancy	-
B181.00	MESOTHELIOMA OF PERITONEUM	MESOTHELIOMA - EXCLUDE	C45
B226.00	MESOTHELIOMA	MESOTHELIOMA - EXCLUDE	C45
B232.00	MESOTHELIOMA OF PLEURA	MESOTHELIOMA - EXCLUDE	C45
B241400	MESOTHELIOMA OF PERICARDIUM	MESOTHELIOMA - EXCLUDE	C45
Byu5000	[X]MESOTHELIOMA OF OTHER SITES	MESOTHELIOMA - EXCLUDE	C45
Byu5011	[X]MESOTHELIOMA OF LUNG	MESOTHELIOMA - EXCLUDE	C45
Byu5100	[X]MESOTHELIOMA, UNSPECIFIED	MESOTHELIOMA - EXCLUDE	C45
BBP1.00	[M]MESOTHELIOMA, MALIGNANT	MESOTHELIOMA - EXCLUDE	C45
BBP5.00	[M]EPITHELIOID MESOTHELIOMA, MALIGNANT	MESOTHELIOMA - EXCLUDE	C45
BBP7.00	[M]MESOTHELIOMA, BIPHASIC TYPE, MALIGNANT	MESOTHELIOMA - EXCLUDE	C45
BBPX.00	[M]MESOTHELIOMA, UNSPECIFIED	MESOTHELIOMA - EXCLUDE	C45
1J0L.00	SUSPECTED MALIGNANT MESOTHELIOMA	MESOTHELIOMA - EXCLUDE	C45
BBP9.00	[M]CYSTIC MESOTHELIOMA	MESOTHELIOMA - EXCLUDE	C45

**Table 10.5: Malignant Neoplasm (Site Unknown) codelist**

Readcode	Description	ICD-10 Code
44a..00	TUMOUR MARKER LEVELS	D48
8A9..00	TUMOUR MARKER MONITORING	D48
B...00	NEOPLASMS	D48
B9...00	NEOPLASMS OF UNCERTAIN BEHAVIOUR	D48
B93y.00	NEOPLASM OF UNCERTAIN BEHAVIOUR OF OTHER SPECIFIED SITES	D48
B9y..00	NEOPLASM OF UNCERTAIN BEHAVIOUR OTHERWISE SPECIFIED	D48
B9z..00	NEOPLASM OF UNCERTAIN BEHAVIOUR NOS	D48
BA...00	UNSPECIFIED NATURE NEOPLASM	D48
BA0..00	NEOPLASM OF UNSPECIFIED NATURE	D48
BA0y.00	NEOPLASM OF UNSPECIFIED NATURE OF OTHER SPECIFIED SITES	D48
BA0z.00	NEOPLASM OF UNSPECIFIED NATURE NOS	D48
BAz..00	NEOPLASM OF UNSPECIFIED NATURE NOS	D48
BB...11	[M]TUMOUR MORPHOLOGY	D48
BB0..00	[M]NEOPLASMS NOS	D48
BB01.00	[M]NEOPLASM, UNCERTAIN WHETHER BENIGN OR MALIGNANT	D48
BB06.00	[M]TUMOUR CELLS, UNCERTAIN WHETHER BENIGN OR MALIGNANT	D48
BB5..00	[M]ADENOMAS AND ADENOCARCINOMAS	D48
BB5y.00	[M]ADENOMA AND ADENOCARCINOMAS	D48
BB5z.00	[M]ADENOMA OR ADENOCARCINOMA NOS	D48
BB80.00	[M]CYSTADENOMA AND CARCINOMA	D48
BBL3.12	[M]MIXED TUMOUR NOS	D48
By...00	NEOPLASMS OTHERWISE SPECIFIED	D48
ByuH.00	[X]NEOPLASMS OF UNCERTAIN AND UNKNOWN BEHAVIOUR	D48
Bz...00	NEOPLASMS NOS	D48



**Table 10.6: Supporting evidence of diagnosis**

<b>Cancer type</b>	<b>Supportive evidence</b>	<b>Description</b>
<b>All cancer types</b>	Non-surgical treatment or support	Chemotherapy, radiotherapy, cancer care (review or plan), terminal illness, palliative care, diagnosis at death, visit to an oncology clinic within 2 years of diagnosis code
<b>Breast</b>	Surgery	Mastectomy; lumpectomy; quadrantectomy; total, partial, or wedge excisions of the breast
	Hormonal Therapy	Tamoxifen; Goserelin; Anastrozole; Exemestane; Letrozole; Aminoglutethimide; Formestane, Testosterone enantate; Toremifene citrate; Trilostane; Fulvestrant
<b>Colorectal</b>	Surgery	Colectomy (transverse, left/right hemi, sigmoid ); colostomy bag; ileostomy; colo-anal anastomosis; colon/rectum excision; colonic polypectomy ; transanal resection; low anterior resection of the rectum; abdominoperineal.
<b>Lung</b>	Surgery	Lobectomy of the lung, pneumonectomy, excision of the lung
<b>Prostate</b>	Surgery	Prostatectomy; orchidectomy; resection of prostate
	Hormonal Therapy	<b>Anti-androgens:</b> Bicalutamide, Flutamide; <b>Pituitary down-regulators:</b> Buserelin, Goserelin acetate; Histrelin, Leuprorelin, Triptorelin acetate; <b>Gonadotrophin releasing hormone blockers:</b> Degarelix, <b>Other:</b> Abiraterone, Cyproterone

## 11 Appendix D: Supplementary materials for Chapter 6

Table 11.1: Demographics for the cohort of any statin users; glaucoma cohort

	a) Any statin vs matched non-users cohort			b) New statin users (Table 6.1)	c) Unmatched cohort of <i>new statin</i> users and new users of glaucoma medications					
	Any Statin		Non-Statins	P-value	<i>New statin</i>		<i>New statin</i>		New Glaucoma	P-value <sup>a</sup>
<b>All Patients</b>	418188		2090482		307646		630814		48310	
<b>Age</b>				1.000						<0.001
30-39	20057	(4.8)	100285	(4.8)	14368	(4.7)	14985	(2.4)	2024	(4.2)
40-49	69201	(16.5)	346005	(16.6)	53773	(17.5)	64685	(10.3)	4178	(8.6)
50-59	117161	(28.0)	585819	(28.0)	91622	(29.8)	148766	(23.6)	7721	(16.0)
60-69	110151	(26.3)	550770	(26.3)	80903	(26.3)	201659	(32.0)	11071	(22.9)
70-79	67828	(16.2)	339076	(16.2)	46576	(15.1)	144918	(23.0)	13443	(27.8)
80+	33790	(8.1)	168526	(8.1)	20404	(6.6)	55801	(8.8)	9873	(20.4)
<b>Sex</b>				0.962						<0.001
Male	224770	(53.7)	1123520	(53.7)	163667	(53.2)	343442	(54.4)	22640	(46.9)
Female	193418	(46.3)	966962	(46.3)	143979	(46.8)	287372	(45.6)	25670	(53.1)
<b>Smoking status</b>				<0.001						<0.001
Non	157388	(37.6)	933690	(44.7)	115053	(37.4)	236121	(37.4)	23668	(49.0)
Current	99191	(23.7)	443407	(21.2)	75132	(24.4)	136589	(21.7)	7365	(15.2)
Ex	155614	(37.2)	608700	(29.1)	115704	(37.6)	255700	(40.5)	16082	(33.3)
Unknown	5995	(1.4)	104685	(5.0)	1757	(0.6)	2404	(0.4)	1195	(2.5)
<b>BMI</b>				<0.001						<0.001
<20	8923	(2.1)	89183	(4.3)	6494	(2.1)	15704	(2.5)	2620	(5.4)

20-25	87506	(20.9)	608397	(29.1)		65977	(21.4)	144417	(22.9)	15556	(32.2)	
>25	289653	(69.3)	1086260	(52.0)		218815	(71.1)	439646	(69.7)	24420	(50.5)	
Unknown	32106	(7.7)	306642	(14.7)		16360	(5.3)	31047	(4.9)	5714	(11.8)	
<b>Alcohol status</b>					<0.001							<0.001
<i>Non</i>	62877	(15.0)	232727	(11.1)		38768	(12.6)	74225	(11.8)	6193	(12.8)	
<i>Ex</i>	15799	(3.8)	53500	(2.6)		12350	(4.0)	29874	(4.7)	1429	(3.0)	
<i>Current</i>	12394	(3.0)	55472	(2.7)		8096	(2.6)	14922	(2.4)	1327	(2.7)	
<i>rare&lt;2u/d</i>	72350	(17.3)	334447	(16.0)		56488	(18.4)	120633	(19.1)	8657	(17.9)	
<i>moderate3-6u/d</i>	186966	(44.7)	950951	(45.5)		146339	(47.6)	300606	(47.7)	22289	(46.1)	
<i>excessive &gt;6u/d</i>	36062	(8.6)	169623	(8.1)		29005	(9.4)	58717	(9.3)	3253	(6.7)	
<i>Unknown</i>	31740	(7.6)	293762	(14.1)		16600	(5.4)	31837	(5.0)	5162	(10.7)	
<b>Diabetes</b>	123753	(29.6)	149989	(7.2)	<0.001	88714	(28.8)	173030	(27.4)	9007	(18.6)	<0.001
<b>CHD</b>	100648	(24.1)	76875	(3.7)	<0.001	69974	(22.7)	125597	(19.9)	4659	(9.6)	<0.001
<b>Heart Failure</b>	16591	(4.0)	31250	(1.5)	<0.001	11877	(3.9)	24867	(3.9)	3064	(6.3)	<0.001
<b>Hypertension</b>	200599	(48.0)	454328	(21.7)	<0.001	148318	(48.2)	328027	(52.0)	19306	(40.0)	<0.001
<b>Hyperlipidaemia</b>	135470	(32.4)	89673	(4.3)	<0.001	99255	(32.3)	190140	(30.1)	4757	(9.8)	<0.001
<b>NSAIDs/Aspirin</b>	134987	(32.3)	457198	(21.9)	<0.001	124978	(40.6)	260032	(41.2)	13894	(28.8)	<0.001
<b>Antihypertensives</b>	184993	(44.2)	490584	(23.5)	<0.001	168613	(54.8)	359534	(57.0)	16540	(34.2)	<0.001
<b>OC</b>	2708	(0.6)	20830	(1.0)	<0.001	2484	(0.8)	2800	(0.4)	491	(1.0)	<0.001
<b>HRT</b>	17643	(4.2)	97491	(4.7)	<0.001	16550	(5.4)	26254	(4.2)	1867	(3.9)	<0.001
<b>Consultations</b>												<0.001
<b>Mean (SD)</b>	8.5	(8.6)	3	(6.1)	<0.001	10.6	(8.9)	10.7	(9.0)	9.2	(9.1)	

<sup>a</sup>P-values (two-sided) were from *t* tests (continuous factor) or chi-square test (categorical factor).

**Table 11.2: Demographics for the Case-control design (time-independent sampling)**

	Control		Case		P-value <sup>a</sup>
<b>All Patients</b>	4773887		106244		
<b>Age</b>					<0.001
30-39	802634	(16.8)	1429	(1.3)	
40-49	1114546	(23.3)	7536	(7.1)	
50-59	966910	(20.3)	17084	(16.1)	
60-69	777310	(16.3)	27524	(25.9)	
70-79	547708	(11.5)	30584	(28.8)	
80+	564779	(11.8)	22087	(20.8)	
<b>Sex</b>					<0.001
Male	2362975	(49.5)	50479	(47.5)	
Female	2410912	(50.5)	55765	(52.5)	
<b>Smoking status</b>					<0.001
Non	2122891	(44.5)	40579	(38.2)	
Current	958166	(20.1)	22029	(20.7)	
Ex	1207647	(25.3)	39250	(36.9)	
Unknown	485183	(10.2)	4386	(4.1)	
<b>BMI</b>					<0.001
<20	232851	(4.9)	6513	(6.1)	
20-25	1355102	(28.4)	33503	(31.5)	
>25	2179559	(45.7)	51516	(48.5)	
Unknown	1006375	(21.1)	14712	(13.8)	
<b>Alcohol status</b>					<0.001
Non	496282	(10.4)	11141	(10.5)	
Ex	204295	(4.3)	4858	(4.6)	
Current	144999	(3.0)	2626	(2.5)	
rare<2u/d	731226	(15.3)	18954	(17.8)	
moderate3-6u/d	1923945	(40.3)	47270	(44.5)	
excessive >6u/d	351913	(7.4)	8068	(7.6)	
Unknown	921227	(19.3)	13327	(12.5)	
<b>Diabetes</b>	527076	(11.0)	16153	(15.2)	<0.001
<b>CHD</b>	277935	(5.8)	12587	(11.8)	<0.001
<b>Heart Failure</b>	132161	(2.8)	5483	(5.2)	<0.001
<b>Hypertension</b>	965782	(20.2)	36328	(34.2)	<0.001
<b>Hyperlipidaemia</b>	340481	(7.1)	11276	(10.6)	<0.001
<b>NSAIDs/Aspirin</b>	1104834	(23.1)	39900	(37.6)	<0.001
<b>Antihypertensives</b>	1280320	(26.8)	46742	(44.0)	<0.001
<b>OC</b>	138682	(2.9)	1206	(1.1)	<0.001
<b>HRT</b>	117649	(2.5)	5468	(5.1)	<0.001
<b>Consultations</b>					<0.001
<b>Mean (SD)</b>	6.8	(9.2)	15.2	(13.5)	

<sup>a</sup>P-values (two-sided) were from t tests (continuous factor) or chi-square test (categorical factor).

**Table 11.3: Immortal time bias weighted relative risk,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Weighted Relative Risk <sup>a</sup> (95% CI)	Weighted $\Delta\beta^b$ (95% CI)
<b>(a) Minimum of 2 statin prescriptions</b>						
<b>Breast Cancer</b>						
Biased	Unexposed	505 031	5.94	6496 (1.3)	1.05	-0.01
	Exposed	117 691	5.88	2154 (1.8)	(0.99, 1.10)	(-0.09, 0.06)
Corrected	Unexposed	502 829	5.85	6411 (1.3)	1.06	
	Exposed	117 691	5.77	2154 (1.8)	(1.00, 1.12)	
<b>Colorectal Cancer</b>						
Biased	Unexposed	1 035 532	5.94	5393 (0.5)	1.03	-0.01
	Exposed	251 556	5.83	1787 (0.7)	(0.97, 1.09)	(-0.09, 0.07)
Corrected	Unexposed	1 030 623	5.85	5342 (0.5)	1.04	
	Exposed	251 556	5.71	1787 (0.7)	(0.99, 1.10)	
<b>Lung Cancer</b>						
Biased	Unexposed	1 035 532	5.94	5382 (0.5)	1.03	-0.01
	Exposed	251 556	5.83	1931 (0.8)	(0.97, 1.10)	(-0.10, 0.07)
Corrected	Unexposed	1 030 623	5.85	5337 (0.5)	1.04	
	Exposed	251 556	5.71	1931 (0.8)	(0.98, 1.11)	
<b>Prostate Cancer</b>						
Biased	Unexposed	530 501	5.94	7178 (1.4)	1.10	-0.01
	Exposed	133 865	5.78	2517 (1.9)	(1.05, 1.15)	(-0.08, 0.06)
Corrected	Unexposed	527 794	5.85	7128 (1.4)	1.11	
	Exposed	133 865	5.66	2517 (1.9)	(1.06, 1.16)	
<b>(b) Minimum of 6 months follow-up</b>						
<b>Breast Cancer</b>						
Biased	Unexposed	488 154	6.04	6426 (1.3)	0.98	-0.06
	Exposed	113 735	6.06	2019 (1.8)	(0.93, 1.03)	(-0.14, 0.01)
Corrected	Unexposed	478 769	5.65	6042 (1.3)	1.04	
	Exposed	113 735	5.56	2019 (1.8)	(0.99, 1.10)	
<b>Colorectal Cancer</b>						
Biased	Unexposed	1 000 777	6.05	5311 (0.5)	0.99	-0.05
	Exposed	242 986	6.02	1725 (0.7)	(0.94, 1.05)	(-0.14, 0.03)
Corrected	Unexposed	980 554	5.65	5084 (0.5)	1.05	
	Exposed	242 986	5.52	1725 (0.7)	(0.99, 1.11)	
<b>Lung Cancer</b>						
Biased	Unexposed	1 000 777	6.05	5318 (0.5)	0.99	-0.06
	Exposed	242 986	6.02	1869 (0.8)	(0.94, 1.06)	(-0.15, 0.03)
Corrected	Unexposed	980 554	5.65	5100 (0.5)	1.06	
	Exposed	242 986	5.52	1869 (0.8)	(0.99, 1.12)	
<b>Prostate Cancer</b>						
Biased	Unexposed	512 623	6.05	7067 (1.4)	1.03	-0.05
	Exposed	129 251	5.98	2403 (1.9)	(0.98, 1.09)	(-0.12, 0.02)
Corrected	Unexposed	501 785	5.66	6796 (1.4)	1.08	
	Exposed	129 251	5.48	2403 (1.9)	(1.03, 1.14)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.4: Protopathic bias weighted relative risk,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Weighted Relative Risk <sup>a</sup> (95%CI)	Weighted $\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Biased (0-day lag)	Exposed	131 581	5.74	2 377 (1.8)	1.03	0.02
	Unexposed	553 656	5.81	6 959 (1.3)	(0.98, 1.08)	(-0.05, 0.10)
Corrected (360-day lag)	Exposed	107 399	5.50	1 888 (1.8)	1.01	
	Unexposed	434 616	5.52	5 512 (1.3)	(0.95, 1.07)	
<b>Colorectal Cancer</b>						
Biased (0-day lag)	Exposed	281 347	5.67	1 948 (0.7)	1.03	-0.01
	Unexposed	1 131 970	5.79	5 707 (0.5)	(0.97, 1.08)	(-0.09, 0.07)
Corrected (360-day lag)	Exposed	231 466	5.45	1 679 (0.7)	1.04	
	Unexposed	895 020	5.51	4 749 (0.5)	(0.98, 1.10)	
<b>Lung Cancer</b>						
Biased (0-day lag)	Exposed	281 347	5.67	2 119 (0.8)	1.04	0.01
	Unexposed	1 131 970	5.79	5 713 (0.5)	(0.98, 1.10)	(-0.08, 0.09)
Corrected (360-day lag)	Exposed	231 466	5.45	1 797 (0.8)	1.03	
	Unexposed	895 020	5.51	4 812 (0.5)	(0.97, 1.10)	
<b>Prostate Cancer</b>						
Biased (0-day lag)	Exposed	149 766	5.60	2 726 (1.8)	1.09	0.02
	Unexposed	578 314	5.77	7 549 (1.3)	(1.04, 1.15)	(-0.05, 0.09)
Corrected (360-day lag)	Exposed	124 067	5.41	2 336 (1.9)	1.07	
	Unexposed	460 404	5.51	6 410 (1.4)	(1.02, 1.12)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup> $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.5: Prevalent user bias weighted relative risk,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Weighted Relative Risk <sup>a</sup> (95%CI)	Weighted $\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Biased (Prevalent user)	Unexposed	812 670	5.31	8 871 (1.1)	0.95	-0.11
	Exposed	169 619	5.28	2 837 (1.7)	(0.90, 0.99)	(-0.18, -0.04)
Corrected (New user)	Unexposed	502 829	5.85	6 411 (1.3)	1.06	
	Exposed	117 691	5.77	2 154 (1.8)	(1.00, 1.12)	
<b>Colorectal Cancer</b>						
Biased (Prevalent user)	Unexposed	1 690 276	5.28	7 312 (0.4)	0.96	-0.09
	Exposed	369 963	5.17	2 475 (0.7)	(0.91, 1.00)	(-0.16, -0.01)
Corrected (New user)	Unexposed	1 030 623	5.85	5 342 (0.5)	1.04	
	Exposed	251 556	5.71	1 787 (0.7)	(0.99, 1.10)	
<b>Lung Cancer</b>						
Biased (Prevalent user)	Unexposed	1 690 276	5.28	7 182 (0.4)	0.96	-0.09
	Exposed	369 963	5.17	2 636 (0.7)	(0.91, 1.01)	(-0.17, 0.00)
Corrected (New user)	Unexposed	1 030 623	5.85	5 337 (0.5)	1.04	
	Exposed	251 556	5.71	1 931 (0.8)	(0.98, 1.11)	
<b>Prostate Cancer</b>						
Biased (Prevalent user)	Unexposed	877 606	5.26	9 453 (1.1)	1.03	-0.08
	Exposed	200 344	5.07	3 380 (1.7)	(0.99, 1.07)	(-0.14, -0.01)
Corrected (New user)	Unexposed	527 794	5.85	7 128 (1.4)	1.11	
	Exposed	133 865	5.66	2 517 (1.9)	(1.06, 1.16)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup> $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.6: Immortal time bias imputed, missing category relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Imputed Relative Risk <sup>a</sup> (95% CI)	Imputed $\Delta\beta^b$ (95% CI)	Missing category Relative Risk <sup>a</sup> (95% CI)	Missing category $\Delta\beta^b$ (95% CI)
<b>(a) Minimum of 2 statin prescriptions</b>								
<b>Breast Cancer</b>								
Biased	Unexposed	642 227	5.67	10 012 (1.6)	1.09	-0.01	1.09	-0.01
	Exposed	128 453	5.72	2 298 (1.8)	(1.03, 1.16)	(-1.00, .07)	(1.03, 1.16)	(-0.09, 0.07)
Corrected	Unexposed	638 649	5.59	9 870 (1.5)	1.10		1.10	
	Exposed	128 453	5.60	2 298 (1.8)	(1.04, 1.17)		(1.04, 1.17)	
<b>Colorectal Cancer</b>								
Biased	Unexposed	1 370 363	5.59	8 905 (0.6)	1.06	-0.02	1.07	-0.02
	Exposed	274 109	5.65	1 904 (0.7)	(1, 1.13)	(-0.11, 0.07)	(1.00, 1.14)	(-0.11, 0.07)
Corrected	Unexposed	1 362 156	5.51	8 790 (0.6)	1.08		1.09	
	Exposed	274 109	5.53	1 904 (0.7)	(1.02, 1.15)		(1.02, 1.16)	
<b>Lung Cancer</b>								
Biased	Unexposed	1 370 363	5.59	9 265 (0.7)	1.08	-0.02	1.08	-0.02
	Exposed	274 109	5.65	2 138 (0.8)	(1.01, 1.15)	(-0.11, 0.07)	(1.01, 1.16)	(-0.11, 0.07)
Corrected	Unexposed	1 362 156	5.51	9 157 (0.7)	1.10		1.10	
	Exposed	274 109	5.53	2 138 (0.8)	(1.03, 1.17)		(1.03, 1.18)	
<b>Prostate Cancer</b>								
Biased	Unexposed	728 136	5.52	11 642 (1.6)	1.15	-0.01	1.14	-0.01
	Exposed	145 656	5.58	2 655 (1.8)	(1.09, 1.21)	(-0.09, 0.07)	(1.08, 1.2)	(-0.09, 0.07)
Corrected	Unexposed	723 507	5.44	11 533 (1.6)	1.16		1.15	
	Exposed	145 656	5.47	2 655 (1.8)	(1.1, 1.23)		(1.09, 1.22)	
<b>(b) Minimum of 6 months follow-up</b>								
<b>Breast Cancer</b>								
Biased	Unexposed	619 081	5.07	9 816 (1.6)	1.02	-0.07	1.02	-0.07
	Exposed	123 823	5.17	2 153 (1.7)	(.96, 1.08)	(-0.15, 0.01)	(0.96, 1.08)	(-0.15, 0.01)
Corrected	Unexposed	604 046	4.76	9 135 (1.5)	1.09		1.09	
	Exposed	123 823	4.75	2 145 (1.7)	(1.03, 1.16)		(1.03, 1.16)	
<b>Colorectal Cancer</b>								
Biased	Unexposed	1 319 720	5.00	8 692 (0.7)	1.02	-0.06	1.03	-0.06
	Exposed	263 976	5.13	1 831 (0.7)		(-0.16, 0.03)	(0.96, 1.09)	(-0.15, 0.03)
Corrected	Unexposed	1 286 440	4.72	8 240 (0.6)	1.09		1.09	
	Exposed	263 976	4.72	1 829 (0.7)			(1.02, 1.17)	
<b>Lung Cancer</b>								
Biased	Unexposed	1 319 720	5.00	9 083 (0.7)	1.04	-0.06	1.05	-0.06
	Exposed	263 976	5.13	2 057 (0.8)		(-0.16, 0.03)	(0.98, 1.12)	(-0.16, 0.03)
Corrected	Unexposed	1 286 440	4.72	8 589 (0.7)	1.11		1.11	
	Exposed	263 976	4.72	2 054 (0.8)			(1.04, 1.19)	
<b>Prostate Cancer</b>								
Biased	Unexposed	700 639	4.95	11 352 (1.6)	1.09	-0.05	1.08	-0.05
	Exposed	140 153	5.09	2 525 (1.8)		(-0.13, 0.03)	(1.02, 1.14)	(-0.13, 0.03)
Corrected	Unexposed	682 394	4.68	10 855 (1.6)	1.14		1.13	
	Exposed	140 153	4.68	2 523 (1.8)			(1.07, 1.2)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates



**Table 11.7: Protopathic bias imputed, missing category relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Imputed Relative Risk <sup>a</sup> (95%CI)	Imputed $\Delta\beta^b$ (95% CI)	Missing category Relative Risk <sup>a</sup> (95% CI)	Missing category $\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>								
Biased (0-day lag)	Unexposed	719 853	5.58	10991 (1.5)	1.11	0.02	1.11	0.02
	Exposed	143 979	5.59	2537 (1.8)	(1.05, 1.17)	(-0.07, .010)	(1.05, 1.17)	(-0.07, 0.10)
Corrected (360-day lag)	Unexposed	553 415	5.34	8304 (1.5)	1.09		1.09	
	Exposed	116 139	5.38	2012 (1.7)	(1.02, 1.16)		(1.02, 1.16)	
<b>Colorectal Cancer</b>								
Biased (0-day lag)	Unexposed	1 538 020	5.48	9674 (0.6)	1.08	-0.01	1.08	-0.01
	Exposed	307 646	5.49	2076 (0.7)	(1.02, 1.15)	(-0.10, 0.08)	(1.02, 1.15)	(-0.10, 0.08)
Corrected (360-day lag)	Unexposed	1 187 314	5.28	7638 (0.6)	1.09		1.09	
	Exposed	249 648	5.34	1779 (0.7)	(1.02, 1.17)		(1.02, 1.17)	
<b>Lung Cancer</b>								
Biased (0-day lag)	Unexposed	1 538 020	5.48	10101 (0.7)	1.10	0.01	1.11	0.01
	Exposed	307 646	5.49	2350 (0.8)	(1.04, 1.18)	(-0.08, 0.11)	(1.04, 1.18)	(-0.08, 0.11)
Corrected (360-day lag)	Unexposed	1 187 314	5.28	8013 (0.7)	1.09		1.10	
	Exposed	249 648	5.34	1968 (0.8)	(1.02, 1.17)		(1.02, 1.18)	
<b>Prostate Cancer</b>								
Biased (0-day lag)	Unexposed	818 167	5.39	12597 (1.5)	1.17	0.02	1.16	0.02
	Exposed	163 667	5.41	2875 (1.8)	(1.11, 1.23)	(-0.05, 0.10)	(1.1, 1.22)	(-0.05, 0.10)
Corrected (360-day lag)	Unexposed	633 899	5.21	10179 (1.6)	1.14		1.13	
	Exposed	133 509	5.29	2446 (1.8)	(1.08, 1.21)		(1.07, 1.20)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.8: Prevalent user bias imputed, missing category relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Imputed Relative Risk <sup>a</sup> (95%CI)	Imputed $\Delta\beta^b$ (95% CI)	Missing category Relative Risk <sup>a</sup> (95% CI)	Missing category $\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>								
Biased (Prevalent user)	Unexposed	966 962	5.08	14 405 (1.5)	1.01	-0.09	1.01	-0.09
	Exposed	193 418	4.90	3 071 (1.6)	(0.96, 1.06)	(-0.17, -0.01)	(0.96, 1.06)	(-0.17, -0.01)
Corrected (New user)	Unexposed	638 649	5.59	9 870 (1.5)	1.10		1.10	
	Exposed	128 453	5.60	2 298 (1.8)	(1.04, 1.17)		(1.04, 1.17)	
<b>Colorectal Cancer</b>								
Biased (Prevalent user)	Unexposed	2 090 482	5.01	12 767 (0.6)	1.03	-0.05	1.03	-0.06
	Exposed	418 188	4.81	2 683 (0.6)	(0.97, 1.08)	(-0.14, 0.03)	(0.97, 1.08)	(-0.14, 0.03)
Corrected (New user)	Unexposed	1 362 156	5.51	8 790 (0.6)	1.08		1.09	
	Exposed	274 109	5.53	1 904 (0.7)	(1.02, 1.15)		(1.02, 1.16)	
<b>Lung Cancer</b>								
Biased (Prevalent user)	Unexposed	2 090 482	5.01	13 058 (0.6)	1.03	-0.07	1.03	-0.07
	Exposed	418 188	4.81	2 978 (0.7)	(0.97, 1.09)	(-0.15, 0.02)	(0.97, 1.09)	(-0.16, 0.02)
Corrected (New user)	Unexposed	1 362 156	5.51	9 157 (0.7)	1.10		1.10	
	Exposed	274 109	5.53	2 138 (0.8)	(1.03, 1.17)		(1.03, 1.18)	
<b>Prostate Cancer</b>								
Biased (Prevalent user)	Unexposed	1 123 520	4.94	16 494 (1.5)	1.08	-0.07	1.08	-0.07
	Exposed	224 770	4.73	3 581 (1.6)	(1.03, 1.14)	(-0.14, 0.00)	(1.03, 1.13)	(-0.14, 0.01)
Corrected (New user)	Unexposed	723 507	5.44	11 533 (1.6)	1.16		1.15	
	Exposed	145 656	5.47	2 655 (1.8)	(1.10, 1.23)		(1.09, 1.22)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.9: Healthy user bias imputed, missing category relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Imputed Relative Risk <sup>a</sup> (95%CI)	Imputed $\Delta\beta^b$ (95% CI)	Missing category Relative Risk <sup>a</sup> (95% CI)	Missing category $\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>								
Biased (non-user comparison group)	Unexposed	638 649	5.59	9 870 (1.5)	1.10	0.08	1.10	0.08
	Exposed	128 453	5.60	2 298 (1.8)	(1.04, 1.17)	(-0.04, 0.20)	(1.04, 1.17)	(-0.04, 0.20)
Corrected (glaucoma medication comparison group)	Unexposed	25 670	4.77	441 (1.7)	1.02		1.02	
	Exposed	277 236	4.82	4 485 (1.6)	(.92, 1.13)		(0.91, 1.13)	
<b>Colorectal Cancer</b>								
Biased (non-user comparison group)	Unexposed	1 362 156	5.51	8 790 (0.6)	1.08	-0.07	1.09	-0.08
	Exposed	274 109	5.53	1 904 (0.7)	(1.02, 1.15)	(-0.20, 0.05)	(1.02, 1.16)	(-0.04, 0.20)
Corrected (glaucoma medication comparison group)	Unexposed	48 310	4.69	410 (0.8)	1.17		1.18	
	Exposed	610 121	4.69	4 565 (0.7)	(1.04, 1.3)		(1.06, 1.32)	
<b>Lung Cancer</b>								
Biased (non-user comparison group)	Unexposed	1 362 156	5.51	9 157 (0.7)	1.10	-0.01	1.10	-0.05
	Exposed	274 109	5.53	2 138 (0.8)	(1.03, 1.17)	(-0.14, 0.12)	(1.03, 1.18)	(-0.18, 0.08)
Corrected (glaucoma medication comparison group)	Unexposed	48 310	4.69	397 (0.8)	1.11		1.16	
	Exposed	610 121	4.69	4 858 (0.8)	(.99, 1.24)		(1.04, 1.3)	
<b>Prostate Cancer</b>								
Biased (non-user comparison group)	Unexposed	723 507	5.44	11 533 (1.6)	1.16	0.06	1.15	0.06
	Exposed	145 656	5.47	2 655 (1.8)	(1.1, 1.23)	(-0.04, 0.17)	(1.09, 1.22)	(-0.04, 0.17)
Corrected (glaucoma medication comparison group)	Unexposed	22 640	4.57	589 (2.6)	1.09		1.08	
	Exposed	332 885	4.59	6 319 (1.9)	(.99, 1.19)		(0.99, 1.19)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.10: Time-window bias imputed, missing category relative risk estimates,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Case status	N	Median Follow-up (years)	Statin user(%)	Imputed Relative <sup>a</sup> Risk (95% CI)	Imputed $\Delta\beta^b$ (95% CI)	Missing Category Relative Risk <sup>a</sup> (95% CI)	Missing category $\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>								
Biased (Time independent sampling)	Control	2 430 456	5.37	350 750	0.80	-0.24	0.76	-0.31
	Case	36 221	5.06	350 750	(0.77, 0.83)	(-0.29, -0.18)	(0.73, 0.79)	(-0.37, -0.26)
Corrected (Risk set sampling)	Control	552 548	6.07	128 880	1.01		1.04	
	Case	32 992	5.68	128 880	(0.97, 1.05)		(1.00, 1.09)	
<b>Colorectal Cancer</b>								
Biased (Time independent sampling)	Control	4 858 163	5.26	780 723	0.89	-0.12	0.84	-0.22
	Case	21 968	5.68	780 723	(0.85, .92)	(-0.18, -0.06)	(0.81, 0.87)	(-0.28, -0.16)
Corrected (Risk set sampling)	Control	565 219	6.04	128 747	1.00		1.04	
	Case	20 321	6.24	128 747	(0.95, 1.04)		(1.00, 1.09)	
<b>Lung Cancer</b>								
Biased (Time independent sampling)	Control	4 857 109	5.27	780 400	0.72	-0.25	0.70	-0.35
	Case	23 022	5.18	780 400	(0.69, 0.75)	(-0.31, -0.19)	(0.67, 0.72)	(-0.42, -0.29)
Corrected (Risk set sampling)	Control	564 461	6.06	128 467	0.92		0.99	
	Case	21 079	5.80	128 467	(0.88, 0.97)		(0.94, 1.04)	
<b>Prostate Cancer</b>								
Biased (Time independent sampling)	Control	2 388 421	5.18	422 280	0.93	-0.04	0.86	-0.15
	Case	25 033	6.19	422 280	(0.90, .96)	(-0.09, 0.01)	(0.83, 0.89)	(-0.20, -0.10)
Corrected (Risk set sampling)	Control	562 275	6.03	126 604	0.97		1.00	
	Case	23 265	6.71	126 604	(0.93, 1.01)		(0.96, 1.04)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.11: Immortal time bias censored relative risk,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Censored Relative Risk <sup>a</sup> (95%CI)	Censored $\Delta\beta^b$ (95% CI)
<b>(a) Minimum of 2 statin prescriptions</b>						
<b>Breast Cancer</b>						
Biased	Unexposed	505 031	4.57	6505 (1.3)	1.15	-0.02
	Exposed	117 691	4.52	1797 (1.5)	(1.07, 1.25)	(-0.12, 0.09)
Corrected	Unexposed	502 829	4.51	6400 (1.3)	1.17	
	Exposed	117 691	4.41	1797 (1.5)	(1.09, 1.27)	
<b>Colorectal Cancer</b>						
Biased	Unexposed	1 035 532	4.45	4913 (0.5)	1.15	-0.02
	Exposed	251 556	4.50	1497 (0.6)	(1.06, 1.25)	(-0.14, 0.10)
Corrected	Unexposed	1 030 623	4.39	4834 (0.5)	1.17	
	Exposed	251 556	4.39	1497 (0.6)	(1.08, 1.28)	
<b>Lung Cancer</b>						
Biased	Unexposed	1 035 532	4.45	4823 (0.5)	1.03	-0.02
	Exposed	251 556	4.50	1561 (0.6)	(0.94, 1.13)	(-0.16, 0.12)
Corrected	Unexposed	1 030 623	4.39	4761 (0.5)	1.05	
	Exposed	251 556	4.39	1561 (0.6)	(0.96, 1.16)	
<b>Prostate Cancer</b>						
Biased	Unexposed	530 501	4.35	6467 (1.2)	1.24	-0.02
	Exposed	133 865	4.49	2135 (1.6)	(1.15, 1.34)	(-0.13, 0.09)
Corrected	Unexposed	527 794	4.29	6389 (1.2)	1.26	
	Exposed	133 865	4.37	2135 (1.6)	(1.17, 1.36)	
<b>(b) Minimum of 6 months follow-up</b>						
<b>Breast Cancer</b>						
Biased	Unexposed	488 154	4.68	6407 (1.3)	1.05	-0.09
	Exposed	113 735	4.72	1662 (1.5)	(0.97, 1.14)	(-0.20, 0.02)
Corrected	Unexposed	478 769	4.40	5926 (1.2)	1.15	
	Exposed	113 735	4.22	1662 (1.5)	(1.06, 1.25)	
<b>Colorectal Cancer</b>						
Biased	Unexposed	1 000 777	4.56	4821 (0.5)	1.08	-0.09
	Exposed	242 986	4.70	1435 (0.6)	(0.99, 1.18)	(-0.21, 0.04)
Corrected	Unexposed	980 554	4.29	4505 (0.5)	1.18	
	Exposed	242 986	4.20	1435 (0.6)	(1.08, 1.29)	
<b>Lung Cancer</b>						
Biased	Unexposed	1 000 777	4.56	4755 (0.5)	0.97	-0.09
	Exposed	242 986	4.70	1499 (0.6)	(0.88, 1.07)	(-0.23, 0.05)
Corrected	Unexposed	980 554	4.29	4464 (0.5)	1.06	
	Exposed	242 986	4.20	1499 (0.6)	(0.96, 1.18)	
<b>Prostate Cancer</b>						
Biased	Unexposed	512 623	4.45	6337 (1.2)	1.13	-0.08
	Exposed	129 251	4.69	2022 (1.6)	(1.05, 1.22)	(-0.19, 0.03)
Corrected	Unexposed	501 785	4.19	5959 (1.2)	1.22	
	Exposed	129 251	4.19	2022 (1.6)	(1.13, 1.32)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.12: Protopathic bias censored relative risk,  $\Delta\beta$  estimates and corresponding 95% confidence intervals**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Censored Relative Risk <sup>a</sup> (95% CI)	Censored $\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Biased (0-day lag)	Unexposed	553 656	4.47	7 131 (1.3)	1.13	0.03
	Exposed	131 581	4.14	1 920 (1.5)	(1.05, 1.22)	(-0.09, 0.14)
Corrected (360-day lag)	Unexposed	434 616	4.35	5 281 (1.2)	1.10	
	Exposed	107 399	4.07	1 437 (1.3)	(1.01, 1.2)	
<b>Colorectal Cancer</b>						
biased (0-day lag)	Unexposed	1 131 970	4.34	5 349 (0.5)	1.13	-0.01
	Exposed	281 347	4.10	1 570 (0.6)	(1.04, 1.23)	(-0.14, 0.11)
Corrected (360-day lag)	Unexposed	895 020	4.23	4 078 (0.5)	1.15	
	Exposed	231 466	4.07	1 307 (0.6)	(1.04, 1.27)	
<b>Lung Cancer</b>						
Biased (0-day lag)	Unexposed	1 131 970	4.34	5 254 (0.5)	1.01	0.04
	Exposed	281 347	4.10	1 632 (0.6)	(0.92, 1.11)	(-0.11, 0.18)
Corrected (360-day lag)	Unexposed	895 020	4.23	4 113 (0.5)	0.98	
	Exposed	231 466	4.07	1 331 (0.6)	(0.87, 1.09)	
<b>Prostate Cancer</b>						
Biased (0-day lag)	Unexposed	578 314	4.22	6 987 (1.2)	1.23	0.03
	Exposed	149 766	4.05	2 241 (1.5)	(1.14, 1.32)	(-0.08, 0.14)
Corrected (360-day lag)	Unexposed	460 404	4.11	5 471 (1.2)	1.19	
	Exposed	124 067	4.05	1 860 (1.5)	(1.09, 1.3)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates

**Table 11.13: Prevalent user bias censored and primary analysis relative risk estimates, corresponding 95% confidence intervals, and percentage difference in risk attributable to prevalent user bias**

Analysis	Statin Exposure	N	Median Follow-up (years)	No. of outcomes (%)	Censored Relative Risk <sup>a</sup> (95%CI)	Censored $\Delta\beta^b$ (95% CI)
<b>Breast Cancer</b>						
Biased (Prevalent user)	Unexposed	812 670	4.13	9 990 (1.2)	0.99	-0.17
	Exposed	169 619	3.82	2 286 (1.3)	(0.93, 1.06)	(-0.27, -0.07)
Corrected (New user)	Unexposed	502 829	4.51	6 400 (1.3)	1.17	
	Exposed	117 691	4.41	1 797 (1.5)	(1.09, 1.27)	
<b>Colorectal Cancer</b>						
Biased (Prevalent user)	Unexposed	1 690 276	4.02	7 640 (0.5)	1.05	-0.12
	Exposed	369 963	3.76	1 994 (0.5)	(0.97, 1.13)	(-0.23, 0.00)
Corrected (New user)	Unexposed	1 030 623	4.39	4 834 (0.5)	1.17	
	Exposed	251 556	4.39	1 497 (0.6)	(1.08, 1.28)	
<b>Lung Cancer</b>						
Biased (Prevalent user)	Unexposed	1 690 276	4.02	7 374 (0.4)	0.96	-0.09
	Exposed	369 963	3.76	2 075 (0.6)	(0.88, 1.04)	(-0.22, 0.03)
Corrected (New user)	Unexposed	1 030 623	4.39	4 761 (0.5)	1.05	
	Exposed	251 556	4.39	1 561 (0.6)	(0.96, 1.16)	
<b>Prostate Cancer</b>						
Biased (Prevalent user)	Unexposed	877 606	3.93	9 993 (1.1)	1.10	-0.14
	Exposed	200 344	3.70	2 741 (1.4)	(1.03, 1.17)	(-0.24, -0.04)
Corrected (New user)	Unexposed	527 794	4.29	6 389 (1.2)	1.26	
	Exposed	133 865	4.37	2 135 (1.6)	(1.17, 1.36)	

<sup>a</sup> Relative risk adjusted for all potential confounders listed in Table 6.2;

<sup>b</sup>  $\Delta\beta$ = Difference between "biased" and "corrected" log relative risk estimates