1

Primary analysis of lipid-regulating agents and dementia outcomes

1.1 Lay summary

Electronic health record (EHR) databases are large collections of medical data, used to manage patient administration and care. Under these systems, whenever a patient attends their GP, their clinical data is recorded in a central database using a standardised coding system. These databases have several advantages over traditional methods of data collection, including the number of people they contain and the length of time for which participants are followed. This is particularly important when studying diseases such as dementia, which may begin to develop in patients long before symptoms are seen.

This analysis makes use of the Clinical Practice Research Datalink (CPRD), which contains the electronic medical records of more than 3 million people from general practices across England. Using this data, the analysis presented in this chapter examined whether treatments which lower cholesterol levels (also known as lipid-regulating agents or LRA) of which statins are a prime example, affect the risk of all-cause dementia and related outcomes (Alzheimer's disease, vascular dementia

and other dementias).

Little evidence for an effect of lipid-regulating agents effect on the risk of Alzheimer's disease was found, with the exception of a slightly increased risk in those prescribed a certain type of lipi-regulatin agent called fibrates. In contrast, I found an increased risk of vascular and other (i.e. non-Alzheimer's) dementia was associated with lipid-regulating agent use.

This increased risk in outcomes with a vascular element (e.g. vascular dementia) is unexpected, and is very likely to be due to the presence of bias in the analysis. This bias, called "confounding by indication", is caused when those who are prescribed a statin are more at risk of vascular dementia for a range of reasons, which makes it appear as if statins are harmful. Despite this limitation, the analysis presented provides an important source of information which will be used in later chapters.

1.2 Introduction

In this Chapter, I present an analysis of a large population-based electronic health record dataset to investigate the relationship between lipid-regulating agent (LRA) use and dementia outcomes. The analysis aims to address important two limitations of the current evidence base as identified by the systematic review presented in Chapter **??**.

Firstly, it explicitly examines vascular dementia as an outcome. The systematic review presented in the previous chapter identified an evidence gap around the effect of lipid-regulating agents on the risk of vascular dementia. As triangulation exercises require as many diverse sources of evidence as possible, this analysis provides a source of information on this outcome.

Secondly, and in a similar vein, the analysis intentionally takes a different analytical approach to that most commonly used to examine the effect of statins on dementia as identified by the systematic review. Specifically, this involved a concerted

effort to address immortal time bias though use of a Cox proportional hazards analysis, incorporating a time-varying treatment indicator.[2] This approach provides a evidence source at risk of a distinct bias due to the alternative analytical strategy. The results from this analysis will be incorporated into the triangulation exercise presented in Chapter **??**.

1.3 Methods

1.3.1 Study protocol

An *a priori* protocol for this study was published,[3] and amendments to this are recorded in Appendix **??**.[4]

1.3.2 Data source

Previously known as the General Practice Research Database (GPRD), the Clinical Practice Research Datalink (CPRD) is a large population-based electronic health record (EHR) database.[5] The database has been collecting primary care data from participating practices across England since 1987.[6,7] It contains the primary care records for more than 10 million primary care patients in England, and is broadly representative of the UK population in terms of age, sex and ethnicity.[5,8]

To avoid the ambiguity of interpreting free-text clinical notes and to allow for easy analysis of the resulting data, the CPRD primarily collects data using a predefined coding system known as Read codes.[9] All clinical events, included clinical test results and diagnoses, can be identified by a specific Read code. The codes use a nested approach (see Table 1.1), with the initial characters defining broad diagnostic topics (e.g. Eu... - Mental and behavioural disorders), while subsequent characters

provide additional information on the specific condition diagnosed (e.g. Eu001 - Dementia in Alzheimer's disease with late onset).

Table 1.1: Example of CPRD Read code hierarchy, showing how "Dementia in Alzheimer's disease with late onset" (*Eu001*) is nested under the top-level of "Mental disorders" (*Eu...*). Broad topics are specified using the initial two alpha-numeric characters of the Read code, while subsequent characters are used to define specific conditions and context.

Level	Read code	Term
1	E....	Mental disorders
2	Eu...	Mental and behavioral disorders
3	Eu0..	Organic mental disorder
4	Eu00.	Dementia in Alzheimer's disease
5	Eu001	Dementia in Alzheimer's disease with late onset

The index events, exposures and outcomes used in this analysis were identified using predetermined code lists, which are available for inspection from the archived repository accompanying this analysis (data/code availability is discussed in Section 1.5.5).

1.3.3 Cohort definition

This analysis included all participants registered at a participating practice between 1 January 1995 and 29 February 2016 who had a flag for "research quality" data. Records pre-dating the 1995 cut-off were included in the original CPRD extract obtained for this analysis. However, these were excluded from the analysis as data quality and reliability are thought to be higher after this date.[10] Additionally, individuals with less than 12 months of continuous records prior to cohort entry were excluded, making the effective start date of the cohort 1 January 1996.

Participants were included in the study cohort if their record contained any of the following index events: a Read code for a diagnosis of hypercholesterolemia or

related condition; a Read code for prescription of a lipid-regulating agent (such as statins); a total cholesterol test result of >4 mmol/L; or an LDL-c test result of >2 mmol/L. These index events allowed me to define a population of participants who were either at risk of hypercholesterolemia, as indicated by the elevated total or LDL cholesterol test results, or had already been diagnosed with it, as indicated by a diagnostic code/related prescription.

The index date for a participant was defined as the date where the first relevant code or test result was recorded on their clinical record, and participants were followed up until the earliest of: (a) an outcome of interest; (b) death; (c) end of follow-up (29 February 2016); (d) last registration date with their GP practice; or (e) the last CPRD collection date for their practice. Participants were ineligble for our cohort if they were less than 40 years of age (as these patients are are less likely to be prescribed a LRA), had less than 12 months of "research quality" data, were simultaneously prescribed more than one lipid-regulating agent (due to the difficult of assigning these to a single exposure group), or were diagnosed with an outcome of interest before or on the date of the index event (i.e. had less than one full day of follow-up).

1.3.4 Exposures

We considered seven lipid-regulating drug classes based on groupings in the British National Formulary (BNF),[11] namely: statins, fibrates, bile acid sequestrants, ezetimibe, nicotinic acid groups, ezetimibe and statin (representing one treatment containing both drugs, rather than the two classes being prescribed concurrently), and omega-3 fatty acid groups.

A participants drug class was assigned based on their first recorded prescription, and any drug switching was ignored in an effort to mimic an intention-to-treat approach. We did however examine how often the initial drug class altered according to one of three criteria:

- **stopped**: defined as the last prescription of the primary class being followed by at least six months of observation;
- **added**: defined as a second drug class being prescribed before the last prescription of the initial class; and
- **switched**: defined as a second drug class being prescribed after the last prescription of the initial class.

1.3.5 Outcomes

We considered five outcomes as part of this analysis: probable Alzheimer's disease, possible Alzheimer's disease, vascular dementia, other dementias, and a composite all-cause dementia outcome. When two or more outcomes were coded in a participant's clinical record, a decision tree was used to differentiate between them (Figure 1.1). The diagnosis date of the outcome was determined by the first record of a relevant code.

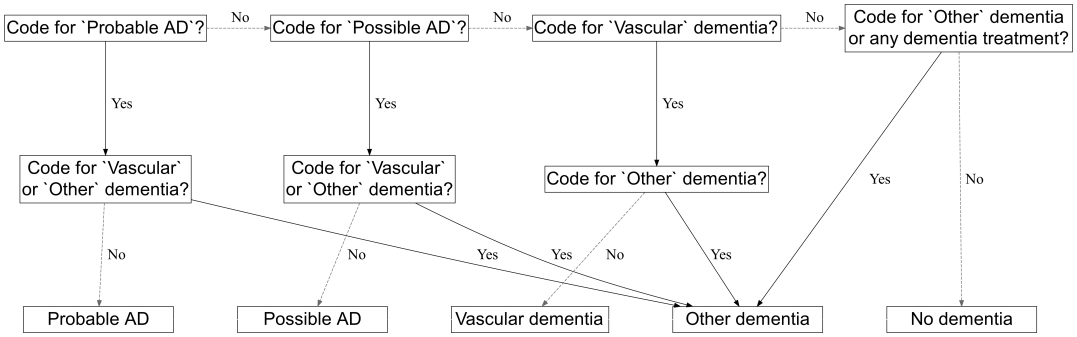

Figure 1.1: Decision tree for assigning dementia subtypes, based on the presence of Read codes in the patient's record. Note that an outcome of "probable" or "Possible" Alzheimer's disease (AD) requires the absence of any vascular outcome codes.

1.3.6 Covariates

A range of additional variables were included in the analysis, intended to address the different distributions of potential confounding variables between those who were prescribed an lipid-regulating agent and those who were not. These are discussed in detail below and summarised in Table 1.2.

Table 1.2: Definition of covariates adjusted for in the Cox PR model. The codelists used to define the majority of these covariates were originally created for use in a previously published analysis.[12] while otherss were built on or adapted from previous published work,**khan2010?**,**taylor2016?**,**wright2017?**.

Covariate	How was the covariate defined?
Previous history of coronary arterial disease	Presence of one or more relevant Read codes on record.
Previous history of coronary bypass surgery	Presence of one or more relevant Read codes on record.
Previous history of cerebrovascular disease (including stroke)	Presence of one or more relevant Read codes on record.
Chronic illness, including cancer and arthritis	Charlson index implemented using Read code lists. (2) Code lists based on those by Taylor et al. (3)
Socioeconomic position	2010 English Index of Multiple Deprivation (IMD) at the twentile level, where 1 represents the least deprived and 20 the most deprived.
Consultation rate	Calculated by dividing the total number of clinic visits by the length of the patient record prior to the index date to give an average annual rate.
Alcohol status	Recorded value (current, former or never).
Smoking status	Most recent of recorded value (current, former or never) or Read code indicating a recorded value. Code lists based on those by Wright et al. (4)
Body Mass Index	Recorded value if available, or a calculated value using the last recorded height and weight measurements. Measurements taken before the age of 25 were excluded to ensure adult measurements were used.
Peripheral arterial disease	Presence of one or more relevant Read codes on record.
Hypertension	Presence of one or more relevant Read codes on record.
Baseline total cholesterol	Continuous value recorded as test result ("enttype==163 & test_data1==3")
Baseline LDL cholesterol	Continuous value recorded as test result ("enttype==177 & test_data1==3")
Chronic kidney disease	Presence of one or more relevant Read codes on record.
Type 1 Diabetes	Presence of one or more relevant Read codes on record.
Type 2 Diabetes	Presence of one or more relevant Read codes on record.

Demographic covariates adjusted for included age and gender. Age was calculated at date of entry into the cohort, and adjusted for via its use as the time axis for the Cox model (see Section 1.3.10). Socioeconomic status was proxied using the Index of Multiple Deprivation (IMD) 2010, which draws on seven domains

(income; employment; education, skills and training; health and disability; crime; barriers to housing and services; living environment) to create an overall deprivation score for each of 32844 statistical geography areas in England. To help preserve patient privacy, IMD score is only available from the CPRD in twentiles, with 1 indicating the least deprived and 20 indicating the most deprived. Smoking and alcohol use was determined at index, and participants were categorised as current, former, or never users of each.

Body mass index (a summary measure calculated as $weight/height^2$), baseline total cholesterol and baseline LDL cholesterol measures were obtained, using the last recorded value prior to the index date. A variable indicating grouped year of entry into the cohort (<2000, 2000-2004, 2005-2009, >2010) was included to allow for changes in prescribing trends across the lifetime of the cohort. To assess healthcare utilisation, I adjusted for the average annual number of consultations between the beginning of a patients data and their entry into the cohort.

Finally, presence of a range of related conditions at baseline were accounted for, including cardiovascular disease, coronary bypass surgery, coronary artery disease, peripheral arterial disease, hypertension, chronic kidney disease, and Type 1 and Type 2 diabetes. In addition to adjusting for these covariates individually, a Charlson co-morbidity index (CCI) score was calculated for each participant. The CCI is a weighted index that uses presence and severity of a number of conditions to enable adjustment for the general health of a participant in terms of their mortality risk.[13] Inclusion of this index allowed me to attempt to adjust for the general health of patients included in the analysis.

Codelists for all covariates can be found in the archived data repository accompanying this analysis (see Section 1.5.5).

1.3.7 Missing data

Missing data are a recognised issue in electronic health records databases,[14] given that they contain administrative data, collected primarily for the purposes of patient management and care rather than academic research.

In this analysis, missing data were handled using a multiple imputation approach.[15] Variables with missing observations were identified for inclusion in the imputation model. Nominal variables with missing values were modelled using multinomial logistic regression, while continuous variables were modelled using linear regression. As per best practice, all variables used in the analytic model, including the outcome, were included in the imputation model.[16] Using the MICE (Multiple Imputation by Chained Equations) command in STATA16, 20 imputed datasets were created.

Missing data was only considered an issue for variables where a numerical test result was expected (e.g. BMI), or where a code existed for the absence of the condition (e.g. categorical smoking status). This approach was necessary, as absence of a code for other treatments or conditions (e.g. statin use or dementia) was assumed to indicate absence of the treatment/condition rather than being considered missing.[14]

1.3.8 Estimation methods

A Cox proportional hazards (PR) model was used to estimate the effect of statins on dementia outcomes. Cox PR models are defined in general terms as:

$$h(t) = h_o(t) \times exp(b_1x_1 + b_2x_2 + ... + b_px_p) \tag{1.1}$$

where:

- t is the survival time;
- $h(t)$ is the hazard function; and

- $x_1, x_2, ..., x_p$ are the covariates which determine the hazard function, while $b_1, b_2, ..., b_p$ are the coefficients for each covariate.
- $h_o(t)$ is the baseline hazard - when all x_i are zero, the $exp()$ function resolves to 1.

A Cox PR model was chosen for this analysis as it inherently accounts for the length of time participants spend in each exposure group. Using this approach, time-at-risk can be properly attributed to the appropriate exposure group, thus mitigating the impact of immortal time bias. This is discussed in detail in the following section.

1.3.9 Immortal time bias and time-varying treatment indicators

Immortal time bias describes two distinct but related types of bias. The first presentation, the selection bias aspect (Panel A, Figure 1.2), occurs when time prior to the exposure is excluded leading to the exposed and control groups being followed up from different time points.[17] Following the example displayed in the figure, the unexposed group are followed from the date of a diagnosis, while the exposed group is followed from date of a prescription. In this scenario, the time between diangosis and prescription for the exposure group is missing, and any events that occur in the exposed group prior to the prescription will be inappropriately excluded from the analysis.

The second presentation of immortal time bias is as a type of misclassification bias (Panel B, Figure 1.2). It occurs when the exposure time prior to the prescription, and any events occurring within it, is inappropriately assigned to the exposed group.

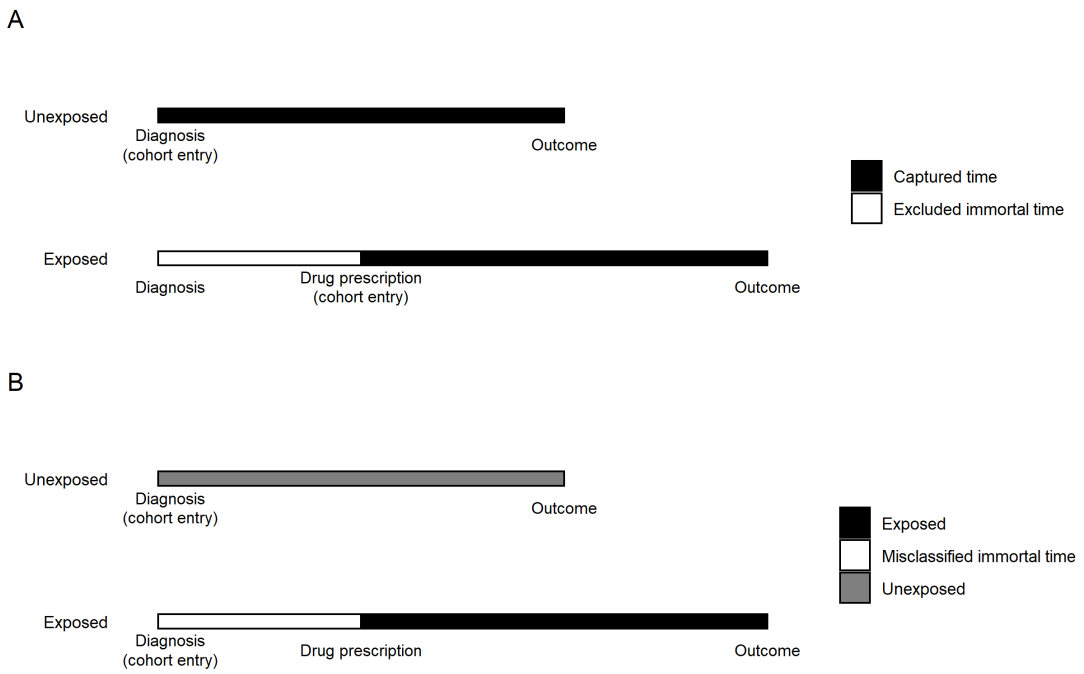

Figure 1.2: Diagram illustrating the two presentations of immortal time bias, as a selection bias (Panel A) and a misclassification bias (Panal B).

This second presentation appears to be common in the existing literature, as several of the studies included in the systematic review presented in Chapter **??** were identified as being at risk of immortal time bias following formal risk of bias assessment using the ROBINS-I tool (see Section **??**). Both presentations of immortal time bias appear in the literature on the relationship of statins and dementia, as identified by the systematic review presented in Chapters **??**/**??**

This analysis attempted to address this issue by following all participants from a common index date (defined as earliest of: (a) date of raised cholesterol test results; (b) hypercholesterolemeia diagnosis; or (c) LRA prescription). Following a recommended approach to addressing the second form of immortal time bias, I employed a time-varying indicator of treatment status to correctly allocate time-at-risk to the exposed and unexposed groups.[17]

Under this approach, all patients start in the unexposed group and contribute time-at-risk until they are prescribed a lipid-regulating agent and move into the exposed group. Note, patients for whom prescription of a lipid-regulating agent was the index event only contribute time to the exposed group (i.e. they enter the cohort and move into the exposed group on the same day).

1.3.10 Time axis

As part of a Cox proportional hazard model, there is the option to use either absolute time in cohort or participants age as the time scale of interest.[18–20] A model using age as the time axis inherently accounts, or adjusts, for participants age as a potential confounder of the exposure-outcome relationship. As such, the main analyses presented all used age as the time axis.

1.3.11 Sensitivity analyses

The primary analysis examined the effect of a lipid-regulating agent on dementia risk, stratified by outcome and drug class. To assess the robustness of the results, a number of sensitivity analyses were performed. These are described in the following sections.

Complete case vs imputed data

Using multiple imputation to handle missing data is an alternative to a "complete case" approach,[21] where participants missing any covariate are dropped from the dataset. As a recommended sensitivity analysis,[22] I preformed and compared the results of both methods, to investigate the impact of multiple imputation on the results.

Control outcomes

In addition to the primary outcomes of interest (described in Section 1.3.5), I extracted data on three additional control outcomes. The inclusion of control outcomes in observational analyses are a useful technique to assess the strength of uncontrolled confounding,[23] and these outcomes are usually class as either "negative" or "positive" outcomes.

Negative outcomes are defined as those without a likely causal path between the exposure and outcome (see Figure 1.3 for a directed acyclic graph, or DAG, describing an ideal negative outcome). Conversely, positive control outcomes are those with a known causal association with the exposure of interest, ideally sourced from large well conducted randomised controlled trials. Positive control outcomes are useful in observational epidemiology, as if the analysis can reproduce a known result for the control outcome, confidence in the result for the outcome of interest is increased. Due to the wealth of data available on statins as a lipid-regulating agent, I chose three control outcomes were chosen in reference to this drug class: back pain (negative control), ischaemic heart disease (positive protective control), and Type 2 diabetes (positive harmful control).

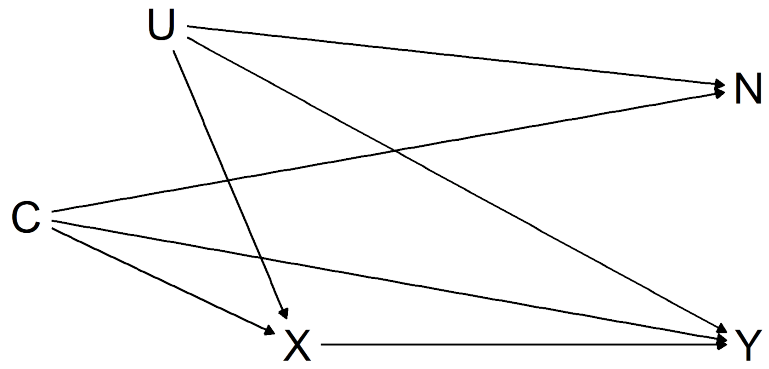

Figure 1.3: Causal diagram (directed acyclic graph) showing relationship between exposure X, outcome Y, confounders (measured C and unmeasured U) and an ideal negative outcome N. Note the absence of any arrow between X and N. In this scenario, any association observed between X and N is due to the presence of uncontrolled confounders U (assuming C has been adjusted for).

Despite observational analyses suggesting a link between statins and muscular pain (as opposed to more serious complications such as myopathy),[24] systematic reviews of the adverse events of statin use[25] and N-of-1 trials explicitly exploring the association of statin use with muscle pain[26] have found little evidence supporting an effect. This suggests that muscular backpain would be suitable for use as a negative control outcome in this analysis. Under this approach, if statin use is found to be associated with muscular backpain in this analysis, this suggests the presence of residual confounding and reduces my confidence in the results for the dementia outcomes.

Similarly, incident ischemic heart disease and Type 2 diabetes were included as a protective and harmful positive control outcome, respectively. The protective effect of lipid-lowering treatment, via statins, on the risk of ischemic heart disease is well-established,[25] while there is growing evidence for an increased risk of Type 2 diabetes with statin use.[25,27,**smit2020?**] Similar to the negative outcome, if the analysis strategy can reproduce these known associations, this will provide evidence that potential confounders have been sufficiently adjusted for.

Impact of additional covariates

To observe the effect of adjusting for additional covariates, I ran two additional models unadjusted except for: (a) age; and (b) age and gender. The results of these models was then compared the results with the fully adjusted model.

Sensitivity cohorts

Two sensitivity cohorts were also created. The first stratified by year of entry into the cohort in an attempt to assess for time period effects. The second removed participants who may have been pregnant (coded as under 55) to assess the robustness of the estimates, as statins are contraindicated in pregnancy.[28]

Statin properties

As detailed in the introduction, the properties of statins may be important in their effect, based on the ability of lipophilic statins to cross the blood brain barrier (see Section **??** in the Introductory Chapter).

To explore whether any observed associations in the statin analysis varied by statin property, a sensitivity analysis was performed, stratifying by statin lipophilicity.

Comparing dementia codelists

As part of an exploratory analysis of the effect of the choice of code lists on the analysis, I created an alternative Alzheimer's disease and non-Alzheimer's dementia outcome using code lists from a published study by Smeeth *et al.*[29]

This previous analysis used a propensity matching approach to estimate the association of statins with a range of outcomes in The Health Improvement Network

(THIN) database, a alternative source of English electronic health records which has substantial overlap with the CPRD.[30] The code lists used in this analysis were obtained through correspondence with the authors of that study, and are available for inspection (see Section 1.5.5).

1.4 Results

1.4.1 Patient characteristics

Of the 3,179,733 participants included in the extract, 1,684,564 met the inclusion criteria (Figure 1.4), with a total follow-up of 10,835,685 patient years at risk.

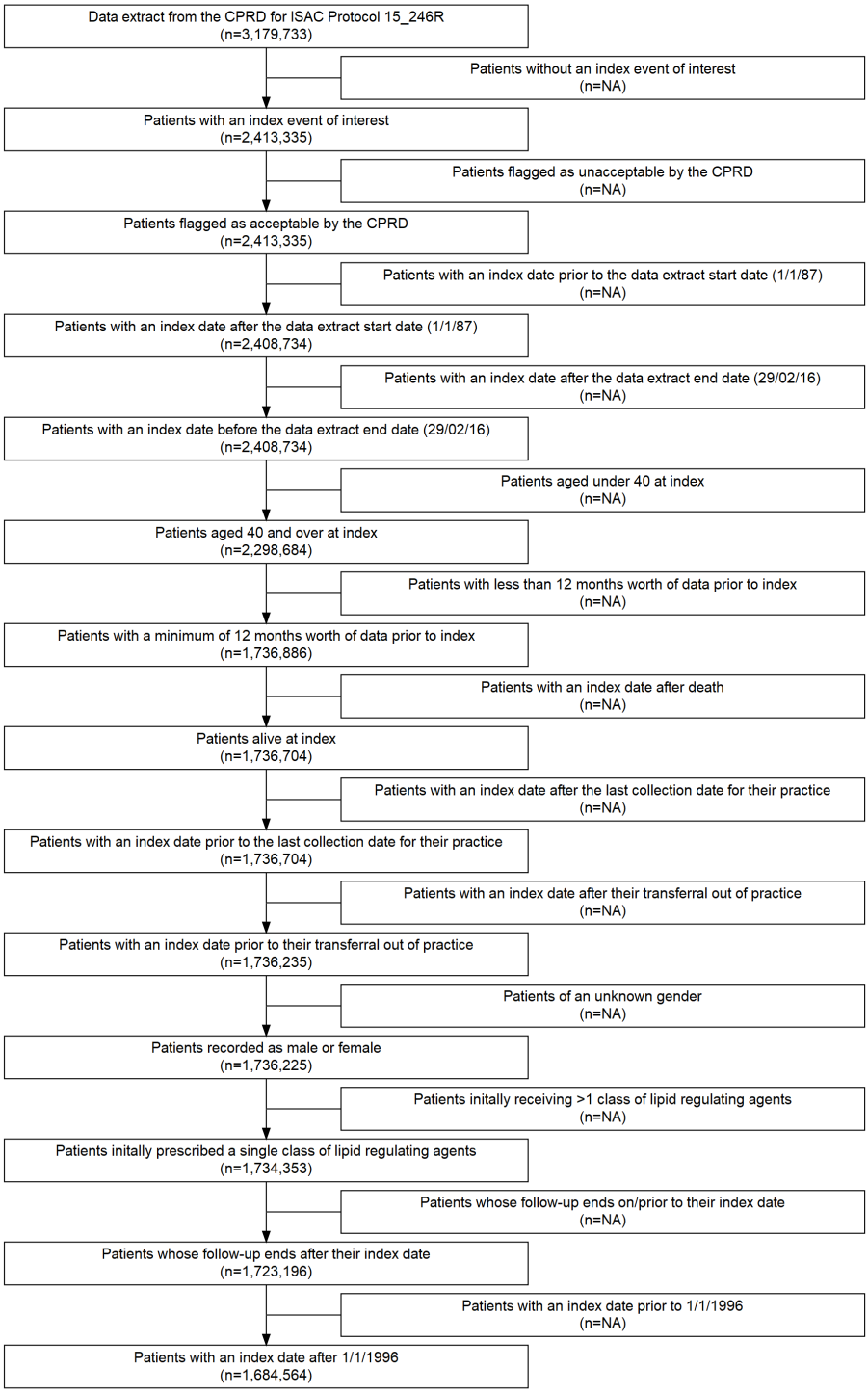

Figure 1.4: Attrition of CPRD participants as the eligibility criteria were applied. The largest cause of attrition was the absence of an index event of interest.

The median participant age at index was 57 years (inter-quartile range (IQR):48-67) and participants were followed up for a median of 5.9 years (IQR:2.7-9.7). During follow-up, an all-cause dementia diagnosis was recorded for 41,830 patients (12,647 probable AD, 9,954 possible AD, 8,466 vascular dementia, 10,763 other dementias).

The number of events, time-at-risk and crude rates for each drug class, tabulated by dementia outcome, are shown in Table 1.3. A substantial majority (98.1%) of participants prescribed a lipid-regulating agent were prescribed a statin. I excluded the "Ezetimibe and statins" (n=127) and "Nicotinic acid groups" (n= 165') classes from subsequent class-based subgroup analyses based on the extremely small number of participants in these groups. Note that the "Ezetimibe and statins" treatment group represent those prescribed a single treatment containing both ezetimibe and statins, rather than those where the two treatments were prescribed concurrently.

Table 1.3: Summary of number of events, years at risk and crude rates per 100,000 participant-years-at-risk stratified by dementia outcome and drug class of interest.

Exposure Group	Any dementia			Possible AD			Probable AD			Vascular dementia			Other dementia		
	Events	PYAR	Rate *	Events	PYAR	Rate *	Events	PYAR	Rate *	Events	PYAR	Rate *	Events	PYAR	Rate *
No LRA (unexposed)	18,608	5,872,717	317	6,368	5,818,047	109	2,637	5,800,964	45	4,813	5,811,594	83	4,790	5,808,285	82
By drug class															
Statins	22,920	4,871,568	470	6,190	4,758,526	130	5,773	4,753,437	121	5,871	4,755,258	123	5,086	4,747,237	107
Omega-3 FAGs	19	8,034	236	4	7,927	50	7	7,950	88	4	7,938	50	4	7,925	50
Fibrates	141	38,003	371	49	37,102	132	21	36,835	57	36	37,001	97	35	36,983	95
Ezetimibe	32	6,604	485	8	6,429	124	7	6,425	109	12	6,444	186	5	6,393	78
BAS	106	36,370	291	28	35,808	78	19	35,726	53	26	35,768	73	33	35,808	92
Ezetimibe + Statins †	0	986	-	0	986	-	0	986	-	0	986	-	0	986	-
NAG	4	1,403	-	0	1,379	-	2	1,391	-	1	1,389	-	1	1,382	-
Total	41,830	10,835,686	386	12,647	10,666,205	119	8,466	10,643,714	80	10,763	10,656,378	101	9,954	10,644,999	94

* Crude rate per 100,000 participant-years-at-risk

† One treatment containing both drugs, rather than the two classes being prescribed concurrently

Abbreviations: AD - Alzheimer's disease; BAS - Bile acid sequestrants; LRA - Lipid regulating agent; NAG - Nicotinic acid groups; Omega-3 FGs - Omega-3 Fatty acid groups; PYAR - Participant-years-at-risk.

Table 1.4: Patient characteristics by drug class. Summary statistics are presented as "% (N)" unless otherwise specified in the variable name.

	Whole Sample	None	Statins	Bile acid sequestrants	Ezetimibe	Fibrates	Omega-3 Fatty Acid Groups
Sample size (N)	1,684,564	1,087,704	585,528	5,396	763	3,889	992
Year of cohort entry (median)	2006	2007	2004	2005	2004	2001	2005
Female	53.0% (893174)	56.2% (610950)	47.1% (276043)	66.4% (3585)	54.5% (416)	38.6% (1500)	52.6% (522)
Age at cohort entry (median)	57	54	62	57	60	58	56
CAD	0.4% (7133)	0.1% (589)	1.1% (6465)	0.1% (6)	0.9% (7)	1.4% (53)	1.3% (13)
CBS	0.3% (5699)	0.1% (682)	0.8% (4926)	0.1% (4)	0.4% (3)	2.0% (78)	0.6% (6)
CVD	2.1% (34899)	1.1% (11619)	3.9% (22977)	1.6% (86)	2.6% (20)	4.4% (170)	1.7% (17)
Charlson (ever > 0)	30.6% (516135)	25.1% (272642)	40.7% (238403)	42.5% (2292)	41.7% (318)	50.8% (1976)	40.4% (401)
IMD-2010 (median)	9	8	9	8	9	10	10
Consultation rate (mean/SD)	5.4 (5.4)	5.0 (5.0)	6.2 (6.1)	8.6 (7.4)	7.4 (6.6)	7.1 (6.2)	8.0 (8.0)
Alcohol (ever)	85.9% (1447151)	86.6% (941648)	84.7% (496110)	82.8% (4468)	84.0% (641)	82.9% (3223)	82.0% (813)
Smoking (ever)	51.1% (861355)	47.1% (511826)	58.6% (343074)	55.2% (2978)	57.5% (439)	60.2% (2341)	53.7% (533)
BMI (mean/SD)	27.0 (5.3)	26.7 (5.2)	27.7 (5.3)	26.8 (5.8)	28.1 (5.7)	29.0 (5.2)	26.9 (5.5)
PAD	0.7% (12613)	0.4% (4039)	1.4% (8424)	0.9% (47)	0.9% (7)	1.9% (75)	1.0% (10)
Hypertension	16.0% (269804)	11.5% (124604)	24.4% (143101)	12.8% (692)	23.9% (182)	25.8% (1002)	15.7% (156)
Total cholesterol (mean/SD)	5.7 (10.1)	5.5 (6.4)	6.2 (15.3)	5.3 (1.3)	7.1 (26.5)	6.4 (5.6)	5.6 (1.6)
LDL cholesterol (mean/SD)	3.6 (4.9)	3.4 (5.3)	4.0 (3.7)	3.1 (1.0)	3.9 (1.1)	3.3 (1.8)	3.2 (1.0)
CKD	0.1% (1295)	0.1% (740)	0.1% (545)	0.1% (6)	0.1% (1)	0.0% (0)	0.3% (3)
Type 1 Diabetes	0.2% (4037)	0.1% (785)	0.5% (3196)	0.3% (14)	1.0% (8)	0.8% (31)	0.1% (1)
Type 2 Diabetes	2.9% (48557)	1.1% (11797)	6.1% (35941)	2.3% (123)	5.4% (41)	15.8% (614)	2.8% (28)

Note: The 'Nicotinic acid groups' (n=165) and 'Ezetimibe and Statins' (n=127) subgroups are not shown, but are included in the whole sample column
Abbreviations: LRA - Lipid regulating agent; IMD - Index of Multiple Deprivation; BMI - Body Mass Index; CAD - Coronary Arterial Disease; CBS - Coronary Bypass Surgery; CVD - Cardiovascular disease; PAD - Peripheral arterial disease; CKD - Chronic Kidney Disease; SD - Standard deviation.

The distribution of baseline characteristics across the remaining seven drug classes can be seen in Table 1.4. Note due to the experimental design, the median year of entry is expected to be later for those not prescribed an LRA, as this exposure group is more likely to include those who entered into the cohort towards the end of study window (as these people had less follow-up time in which to be prescribed an LRA).

The stopping, addition and switching of drug classes was common across all drug classes (Table 1.5).

Table 1.5: Participants who stopped, switched or added treatments by initial treatment type.

	Whole Sample	**Statins**	**Bile acid seques-trants**	**Ezetimibe**	**Ezetimibe & Statins**	**Fibrates**	**Nicotinic acid groups**	**Omega-3 Fatty Acid Groups**
Stopped	6.9% (115899)	19.1% (111798)	56.1% (3028)	19.7% (150)	12.6% (16)	12.3% (478)	44.8% (74)	35.8% (355)
Added	1.6% (27441)	4.4% (25990)	3.6% (192)	19.0% (145)	3.9% (5)	21.6% (841)	3.6% (6)	26.4% (262)
Switched	0.9% (14935)	2.0% (11996)	11.3% (612)	34.6% (264)	64.6% (82)	44.0% (1713)	45.5% (75)	19.5% (193)

Definitions:
Stopped - last prescription of the primary drug class followed by at least six months of observation with no further prescriptions; Added - second drug class prescribed before the last prescription of the initial class; Switched - second drug class being prescribed after the last prescription of the initial class.

1.4.2 Missing data

Full covariate information was available for 450,234 participants (26.7%). Six key variables had some missing data: IMD 2010 score was missing for 625,788 participants (37.1%), because it is only recorded for English practices; alcohol status was missing for 269,526 participants (16%); smoking status was missing for 84,424 participants (5%); BMI, or a calculated BMI from height and weight measurements, was missing for 266,672 participants (15.8%); baseline total cholesterol was missing

for 119,675 participants (7.1%); and baseline LDL cholesterol was missing for 787,289 participants (46.7%).

1.4.3 Primary analysis

The results of the primary analysis using the fully adjusted Cox proportional hazards model with participant age as the time scale are presented for each drug/outcome combination in Figure 1.5.

For each outcome, the overall "Any drug" estimate was driven by the statin subgroup, based on its large size relative to the other drug classes.

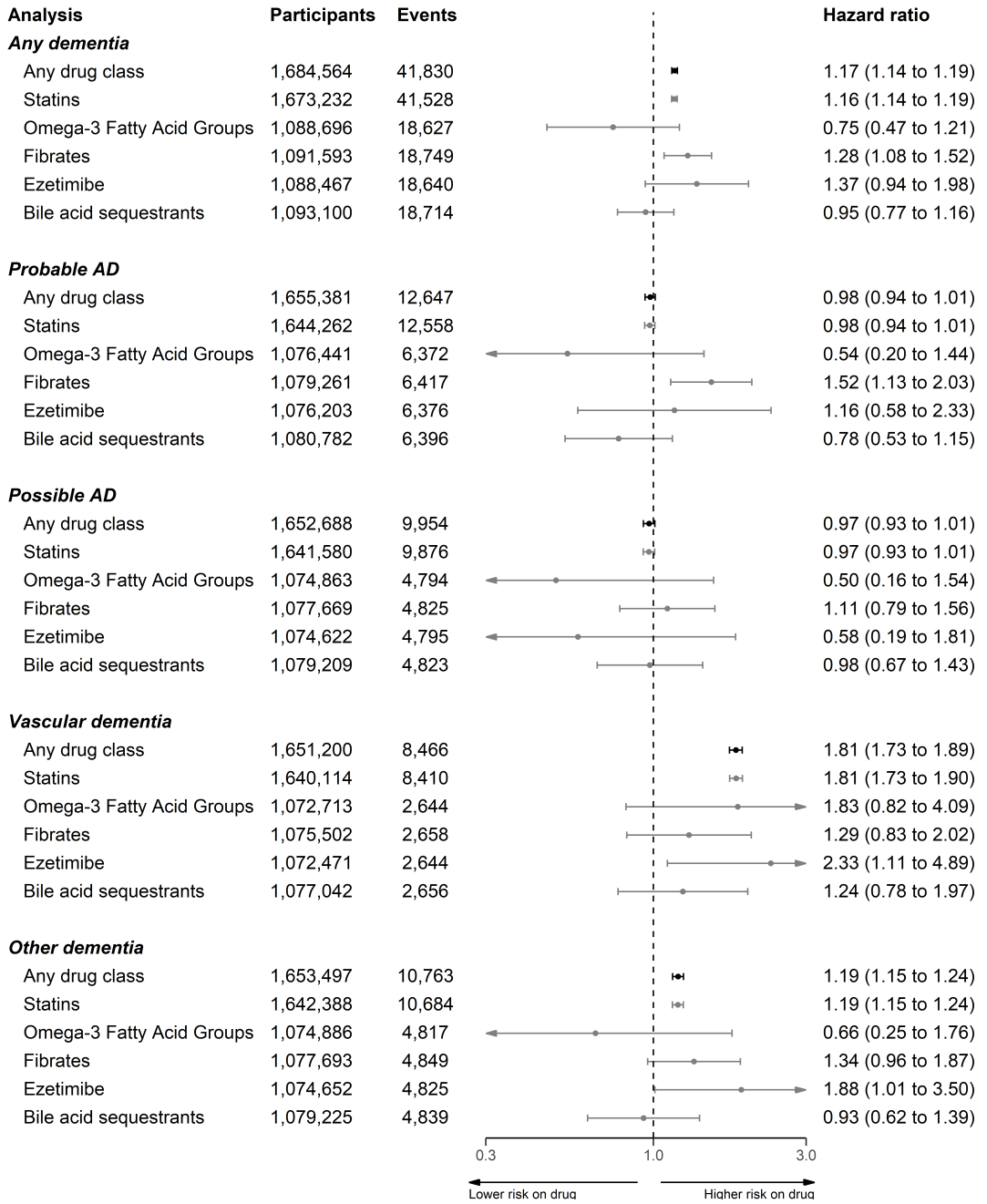

Figure 1.5: Results from primary analyses of CPRD data using the fully adjusted model and participant age as the time scale.

Alzheimer's disease

My results show litte evidence was found for an effect of lipid-regulating agents on probable (HR:0.98, 95%CI:0.94-1.01) and possible (HR:0.97, 95%CI:0.93-1.01)

Alzheimer's disease when compared to no treatment, with the sole exception of fibrates on probable Alzheimer's disease (HR:1.52, 95%CI:1.13-2.03).

Non-Alzheimer's disease dementias

In contrast to the findings for Alzheimer's disease outcomes, lipid-regulating agents were associated with an increased risk of a subsequent diagnosis of vascular dementia (HR:1.81, 95%CI:1.73-1.89) or other dementias (HR:1.19, 95%CI:1.15-1.24). Again this effect was driven mainly by the statin subgroup, but there was some evidence that ezetimibe was associated with an increased risk of vascular (HR:2.33, 95%CI:1.11-4.89) and other (HR:1.88, 95%CI:1.01-3.5) dementia.

All-cause dementia

For the composite all-cause dementia outcome, we found treatment with a lipid-regulating agent was associated with a slightly increased risk (HR:1.17, 95%CI:1.14-1.19), which lies between the associations for the Alzheimer and non-Alzheimer dementia outcomes as would be expected. There was also some evidence that fibrates were associated with increased risk of all-cause dementia (HR:1.28, 95%CI:1.08-1.52).

1.4.4 Sensitivity analyses

The results of the series of sensitivity analyses performed are described in the following sections.

Complete case versus imputed data

In almost all cases, the use of imputed data resulted in a marginal attenuation of the effects observed when using a complete cases analysis. It should be noted that due to the large amount of missing data (e.g. 787,289 participants (46.7%)

were missing a baseline LDL cholesterol measure), the number of participants included in the complete case analysis was substantially smaller than that included when using imputed data. In this case, though the overall position of the effect estimates does not change substantially when using the imputed dataset, there is a noticeable gain in power.[15]

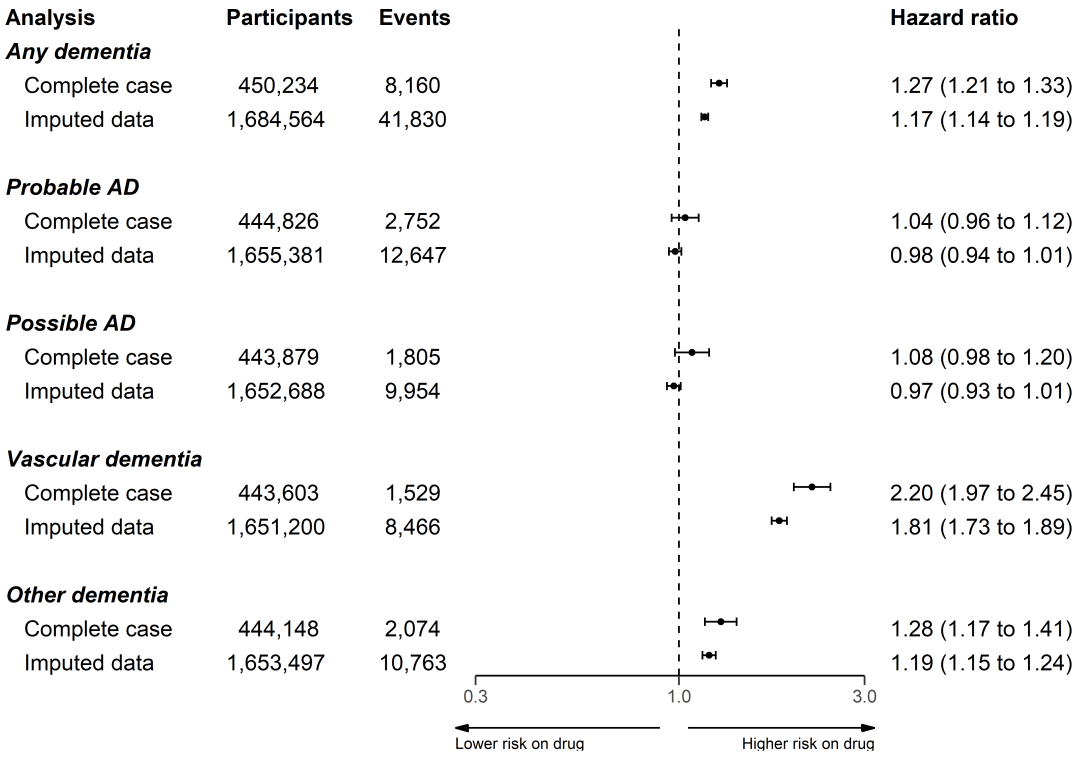

Figure 1.6: Comparison of analyses using the complete case versus imputed cohorts.

Control outcomes

Following the primary analysis, the fully adjusted model was used to estimate the effect of treatment with a statin on the two control outcomes of back pain (negative) and ischemic heart disease (positive). The results of this analysis are presented in Figure 1.7.

For the negative control, there was some evidence that treatment with a statin was associated with an increased risk of back pain (HR: 1.04, 95%CI: 1.03-1.05), suggesting there may be some residual confounding. However, statin prescription was also associated with a substantially increased risk of ischemic heart disease (HR: 1.62, 95%CI: 1.59-1.64) and Type 2 diabetes (HR: 1.50, 95%CI: 1.48-1.51).

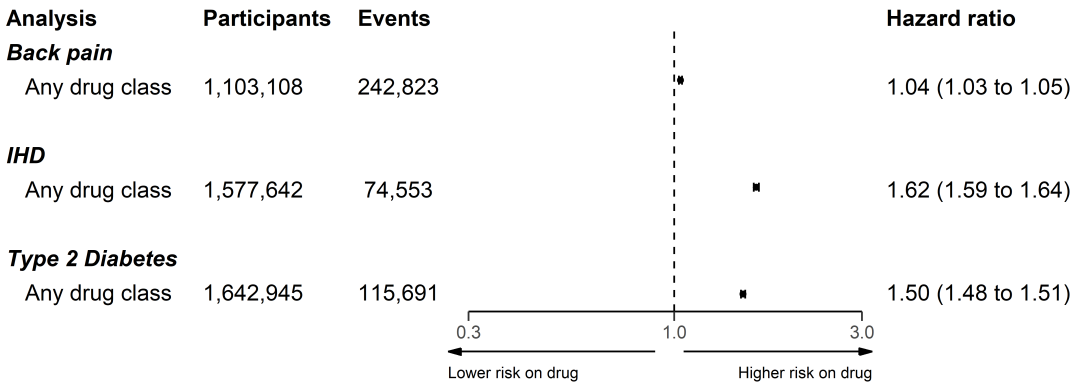

Figure 1.7: Results of control outcome analysis.

Impact of additional covariates

The results of three models adjusted for age only, age and sex, and the full covariates respectively, are presented in Figure 1.8. These models were used to estimate the impact of adjustment for additional covariates. Note that obtaining an completely unadjusted model is not possible, as age was used in the Cox model as the time scale.

Adjustment for additional covariates beyond age and sex had a limited impact on the observed effect estimates, with the exception of the probable AD outcome. In this case, adjustment for the full set of covariates attenuated to the null the protective effect observed when adjusting only for age and sex.

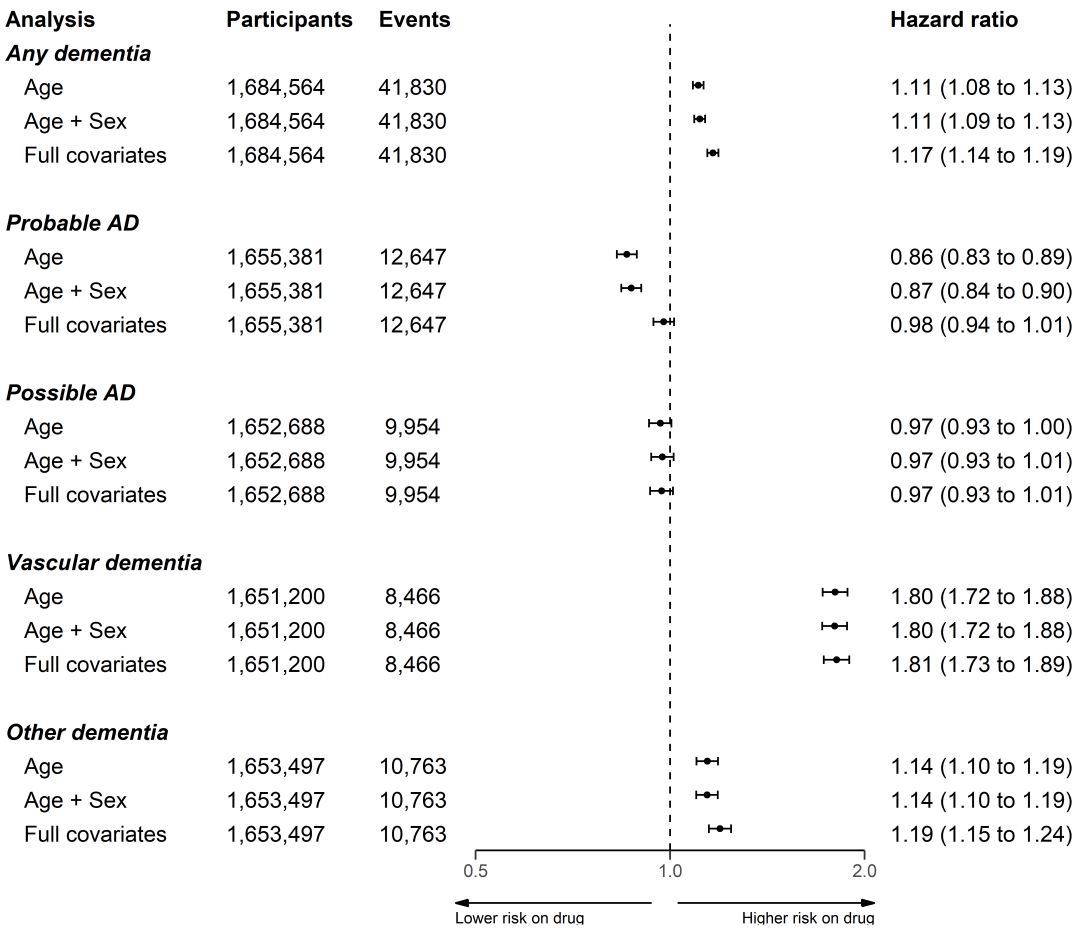

Figure 1.8: Results of three models adjusting for a different set of covariates.

Sensitivity cohorts: Entry year

When stratifying based on year of entry to the cohort, I observed no variation in risk by time period in any subgroup except for probable Alzheimer's disease (Figure 1.9).

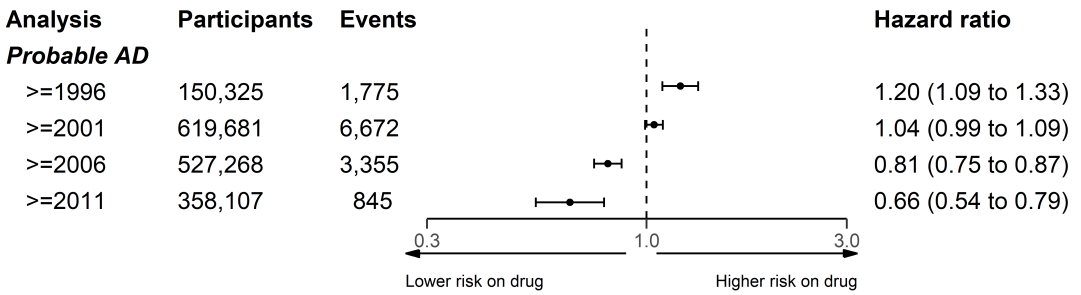

Figure 1.9: Analysis of any lipid-regulating agent on probable AD outcome, stratified by grouped year of cohort entry.

On the assumption that this variation could be caused by changes in the frequency of codes used to define probable AD in the cohort, I performed a *post-hoc* investigation of the frequency of each diagnoses stratified by year of entry (Table 1.6). While the frequency of outcomes declines in more recent strata, likely due to the limited follow-up inherent to these groups, this decline in frequency is relatively constant across the dementia subtypes.

Table 1.6: Frequency of diagnoses by grouped year of cohort entry

Year of cohort entry	No dementia	Probable AD	Possible AD	Vascular dementia	Other dementia	Total
>=1996	148550 (95.9%)	1775 (1.1%)	1677 (1.1%)	1345 (0.9%)	1585 (1.0%)	154932
>=2001	613009 (96.3%)	6672 (1.0%)	5711 (0.9%)	4857 (0.8%)	6073 (1.0%)	636322
>=2006	523913 (98.1%)	3355 (0.6%)	2169 (0.4%)	1890 (0.4%)	2506 (0.5%)	533833
>=2011	357262 (99.4%)	845 (0.2%)	397 (0.1%)	374 (0.1%)	599 (0.2%)	359477
Total	1642734 (97.5%)	12647 (0.8%)	9954 (0.6%)	8466 (0.5%)	10763 (0.6%)	1684564

Sensitivity cohorts: Pregnancy

In the second sensitivity cohort, removing participants aged 55 and under at index from the analysis had minimal effect on the effect estimates (Figure 1.10).

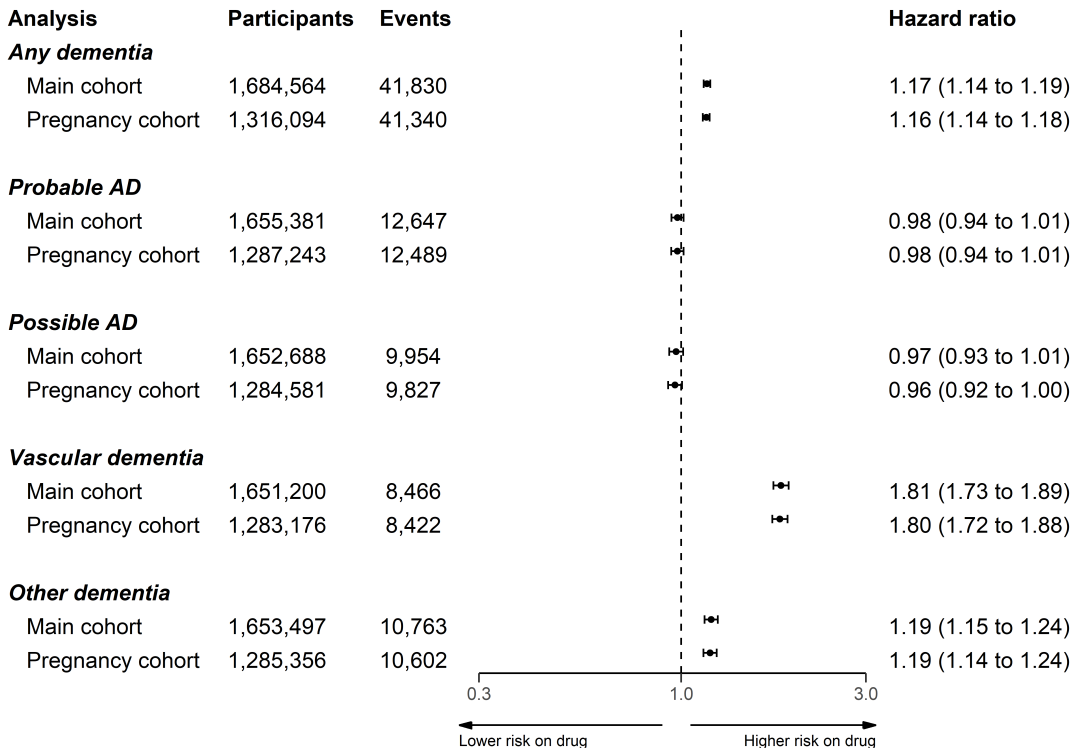

Figure 1.10: Comparison of analysis using main cohort and a cohort with potentially pregnany women removed.

Statin properties

In the cohort, statins with lipophilic properties were much more frequently prescribed than hydrophilic statins (Table 1.7). Additionally, there is evidence for a increasing tendency to favour hydrophilic statins in recent years, as the proportion of lipophilic statins prescribed fell from 18.2% in 1996-2000 to <1% in 2011-2016.

Table 1.7: Summary of statin properties (lipophilicity vs hydrophilicity) by grouped year of prescription.

Prescription Year Group	Hydrophilic	Lipophilic	Total
>=1996	7037 (18.2%)	31531 (81.8%)	38568
>=2001	21427 (10.3%)	187018 (89.7%)	208445
>=2006	3566 (1.6%)	217726 (98.4%)	221292
>=2011	1115 (0.9%)	119035 (99.1%)	120150

When stratifying by statin properties, hydrophilic statins were less harmful in the any, vascular and other dementias outcomes compared to lipophilic statins (Figure 1.11). Additionally, in the AD outcomes, hydrophilic statins were associated with a small reduction in risk, compared to the weak evidence for an effect for lipophilic statins.

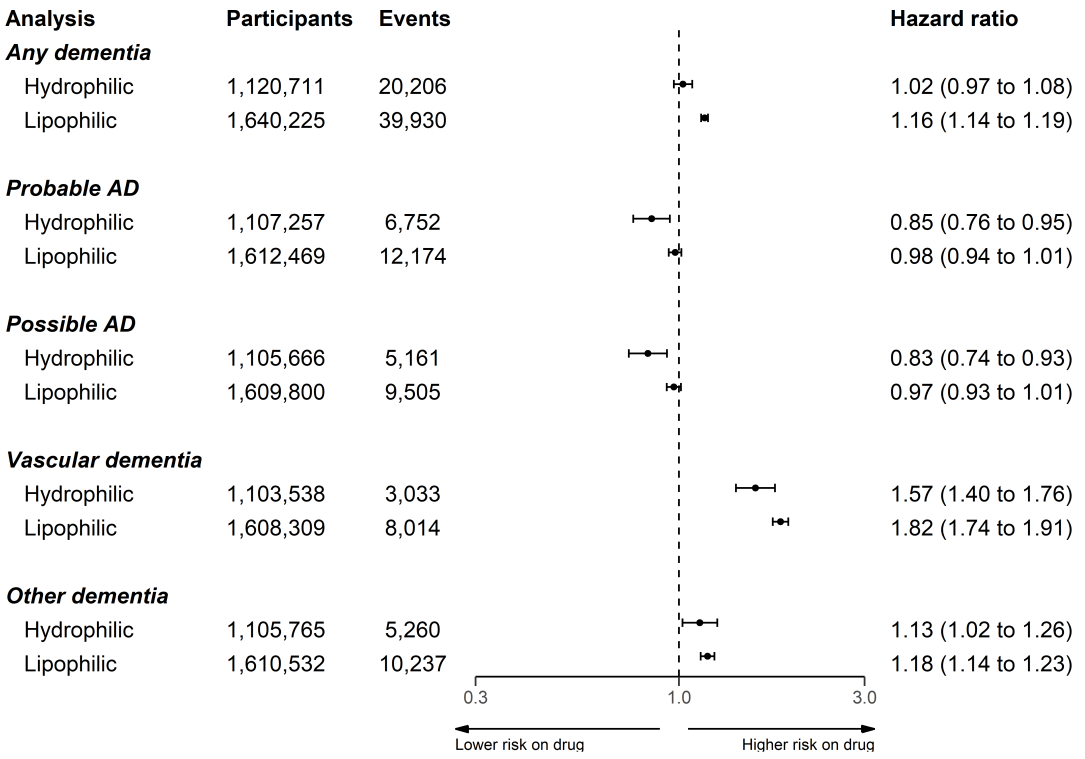

Figure 1.11: Analysis stratified by statin properties (hydrophilic vs lipophilic)

Impact of dementia code lists

When using the Smeeth *et al.* code lists to define dementia outcomes, effect estimates of HR: 1.19 (95%CI: 1.07-1.32) and HR: 1.33 (95%CI: 1.26-1.42) were obtained for the Alzheimer's disease and non-Alzheimer's ("other") dementia outcomes, respectively.

The authors found evidence for a protective effect of statin use on all-cause dementia (HR: 0.81, 95%CI: 0.69-0.96) and non-AD dementia (HR: 0.82, 95%CI: 0.69-0.97), but little evidence of an effect on AD (HR: 0.81, 95%CI: 0.49-1.35).

However, comparison of the results using the two sets of code lists was deemed less useful following a comparison of the codes used. While all of the codes used to define Alzheimer's in the Smeeth paper are included in the probable Alzheimer's code-list (see Figure 1.12), I included several additional codes used to define this outcome (including, for example, "Eu00013: [X]AD disease type 2"). Additionally, several of the codes used to define "Possible Alzheimer's" in this analysis are included in the "Other dementia" code list used by Smeeth.

This analysis serves to illustrative the importance of the code lists chosen to define the outcomes of interest, particularly if they are used to define competing outcomes (e.g. AD vs non-AD dementia).

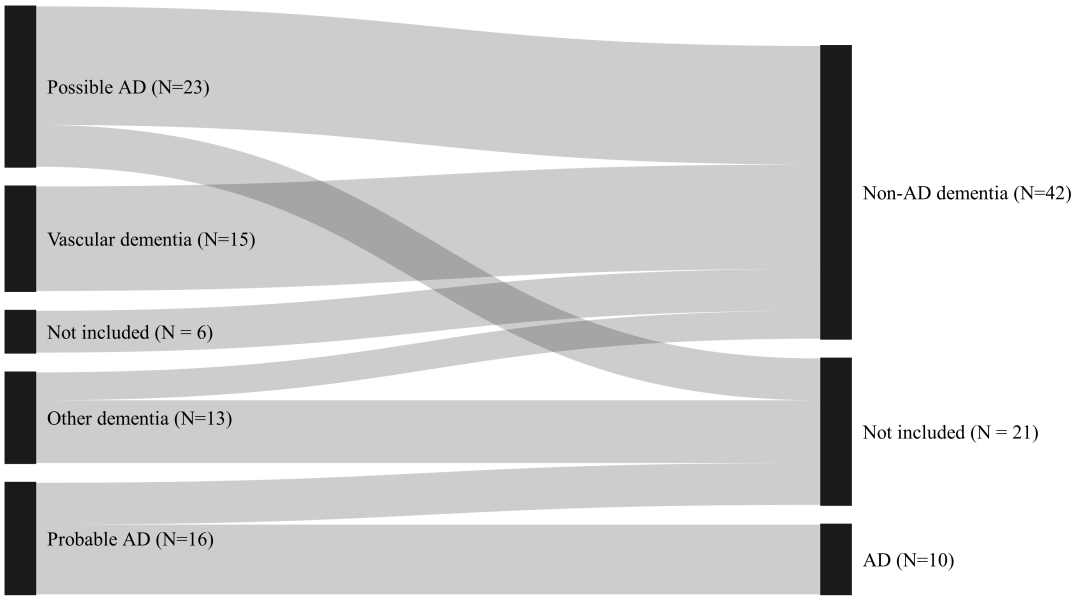

Figure 1.12: Sankey diagram comparing the codes used in this analysis with those used in the Smeeth *et al* paper.[29] The outcomes and number of codes contributing to each are presented (the Smeeth *et al* outcomes are on the right hand side of the figure). The joining lines showing the overlap between the categories in the two analyses.

1.5 Discussion

1.5.1 Summary of findings

Lipid-regulating agents showed little evidence of an association with probable and possible Alzheimer's disease when compared to no treatment, but were associated with increased risk of an all-cause dementia, vascular dementia and other dementias diagnosis. The estimate observed in each case was driven by the statin subgroup, which included a substantial majority of participants. For the other drug classes, no association was found with any outcome, with two exceptions being that ezetimibe was associated with increased risk of vascular and other dementias, while fibrates were associated with increase risk of all-cause dementia and probable Alzheimer's disease.

The effect estimates were robust to the exclusion of potentially pregnant participants, and for all outcomes except probable AD, no variation across grouped year of entry was observed. When looking at the statin subgroup alone, statin properties appeared to have a modifying effect, with hydrophilic statins being less harmful in the any, vascular and other dementias outcomes compared to lipophilic statins.

1.5.2 Comparison with existing literature

[Feedback requested: How much should I be cross-referencing with the systematic review at this stage? Or should I keep it self-contained and only refer to ?]

Statins and all-cause dementia

This conflicts with the findings of our analysis, where statin use was associated with an increased risk of all-cause dementia (HR:1.17, 95%CI:1.14-1.19). Some of the included studies in the meta-analysis specifically exclude vascular dementia from the definition of all-cause dementia,[31] which may limit the ability for comparison with our findings for the all-cause dementia outcome.

Additionally, a previous analysis of the THIN EHR database using a propensity-score matched analysis found a protective effect of statins on all-cause dementia (HR:0.81, 95%CI:0.69-0.96)[29].

Statins and Alzheimer's disease

Our results are broadly in line with the findings of two distinct approaches examining the effect of statin treatment on subsequent Alzheimer's disease. No randomized trials of statins for the prevention of Alzheimer's disease have been reported, but a recent meta-analysis of 20 observational studies found statins were associated with a reduced risk of Alzheimer's disease (RR 0.69, 95% CI 0.60–0.80) with stronger evidence than observed in our analysis.[32] This review included case-control studies

and analyses likely to be at risk of immortal time bias, which may account for the discrepancy with our findings. Additionally, a recent Mendelian randomization study examining the effect of genetic inhibition of HMGCR on Alzheimer's disease (a genetic proxy for statin treatment) provided equivocal evidence (OR: 0.91, 95%CI: 0.63-1.31) but was consistent with our results.[33]

Our additional analyses stratified by statin properties found little evidence of differences in associations of lipophilic and hydrophilic statins and incidence of Alzheimer's disease, consistent with a recent meta-analysis of observational studies.[34]

Statins and vascular/other dementia

Far fewer studies have tested the association between lipid-regulating agents and vascular dementia or other dementias. A recent review found four observational studies examining the association of statins and vascular dementia found limited evidence for an effect (RR:0.93, 95% CI 0.74–1.16).[32] This contrasts with the harmful association found in our analysis (HR:1.81, 95%CI:1.73-1.89). When stratifying by lipid properties, lipophilic statins were more harmful than hydrophilic statins in vascular dementia, potentially due to their ability to cross the blood brain barrier.

Other drug classes

Apart from statins, few studies examining a lipid-regulating agent have been reported (Chapter **??**/**??**).

One of the few classes for which a evidence was available were fibrates, which found little evidence of an association with all-cause dementia was identified,[35] inconsistent with our finding that patients prescribed fibrates had higher all-cause dementia risk than those prescribed other lipid lowering agents.

A previous Mendelian randomization study found little evidence that genetic variants that proxy for ezetimibe affect risk of Alzheimer's disease (OR: 1.17, 95%CI: 0.73-

1.87),[33] consistent with our findings. Note that this study was published in 2020 and so was not included in the review.

1.5.3 Interpretation of results

This section will expand on a potential explanation for the observed results detailed above. However, as the comparison of evidence across different sources is the aim of the triangulation exercise presented in later chapters, the section will not provide a detailed comparison with other published literature, except where needed to illustrate a methodological point. For a comparison of the result presented above with those from the systematic review (Chapter **??**) and the individual patient data analysis (Chapter **??**), see Chapter **??**.

Confounding by indication

A likely explanation for the observed increased risk of vascular and other dementias with lipid-regulating agent use, and an important limitation of this analysis, is residual confounding by indication. While the term has been used to describe different source of bias in epidemiological analyses,[36] "confounding by indication" is used here to described the role of risk factors that both prompt treatment (in this case statins) and increase the risk of the outcome (in this case vascular dementia), thus causing a distorted positive association between the treatment and outcome (see Figure 1.13). In causal inference nomenclature, statins and dementia are said to be *d*-connected as there is an open "backdoor" path between them via the uncontrolled confounders.[37]

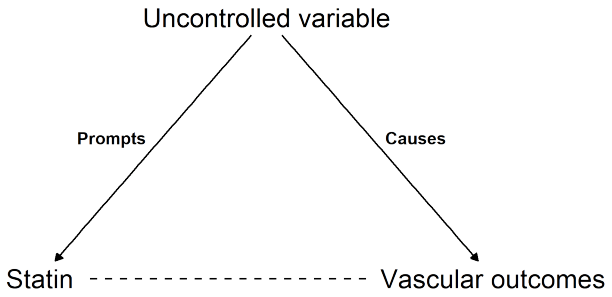

Figure 1.13: Causal diagram (directed acyclic graph) illustrating confounding by indication.

In the case of this analysis, the confounding variable (or, more likely, variables) would prompt prescription of statins but also represent a vascular risk factor that contributes to the development of the vascular dementia.

Supporting evidence for this interpretation comes from a variety of sources, including the results of the control outcome analyses. The slight harmful effect for the backpain outcome is substantially smaller that that observed for the ischemic heart disease outcome, indicating that the majority of the uncontrolled confounding is likely related to vascular factors. This is supported by the increasingly harmful effect moving from probable/Possible AD to other dementias to vascular dementia, indicating that the confounding by indication likely increases as the proportion of vascular outcomes increase. This is supported by the decision tree for assigning outcomes in the presence of greater than one dementia code, where the Alzheimer's disease outcomes require a "pure" condition, and the presence of any vascular or other dementias codes excludes participants from this group (Figure 1.1).

A review of other available literature suggests that this observation (a harmful effect of lipid-regulating agents on vascular-related outcome) is not unusual. Using a conventional epidemiological technique, a previous analysis also found an increased risk of coronary heart disease (analogous to ischemic heart disease) in those taking statins (HR: 1.31, 95%CI: 1.04-1.66).[38] Controlling for confounding by indication in that study through the use of a trial emulation analysis gave an estimate of 0.89

(95%CI: 0.73-1.09), a comparable though less precise estimate to that observed in RCTs of statin use (0.73, 95%CI: 0.67-0.80).**taylor2013?**

Conditioning entry into the study on being either "at-risk" or already diagnosed with hypercholesterolemia was employed in a pre-emptive attempt to mitigate confounding by indication, but evidence from the control outcomes suggests this was unsuccessful.

On a more general note, the unexpected increase in vascular dementia risk with statin use is particularly interesting given the absence of vascular dementia in the published literature, as highlighted in the previous chapter (see Section **??**). It is possible that previous research encountered similar methodological issues to this analysis, and via a publication bias mechanism, the results from these analyses did not make it into the evidence base.

Time period effect in probable Alzheimer's disease

The time period effect observed for the Alzheimer's disease outcome, where statins appear more protective in participants who entered the cohort at a later date, is an interesting result. An expected cause was an change in how different dementia sub-types were diagnosed over time, but no corresponding inverse trend in any of the other dementia outcomes. For example, if the observed trend was due to increasing differential misclassification with vascular dementia, then we would expect vascular dementia to appear increasingly harmful in more recent time periods which is not the case.

Statin properties

The observation that hydrophilic statins appear more protective than their lipophilic counterparts, runs counter to the theory that lipophilic drugs are able to cross the blood brain barrier. However, it does agree with previous findings for this comparison.

Additionaly, the relative precision of the two groups is expected, as the two most commonly prescribed statins in the UK are lipophilic (simvastatin and atorvastatin).

An initial interpretation for the discrepancy was that lipophilic statins may as a class be stronger, and so are prescribed to patients with a higher underlying vascular load, leading to increased confounding by indication in this group. However, the statin with the strongest lipid lowering effect available via the NHS, rosuvastatin, is hydrophobic.

however, two of the most widely used

[Yoav: are there are particular indications/contraindications for statins with different properties?]

- Blood brain barrier
-
-

1.5.4 Strengths and limitations

The primary strength of this analysis compared to others available in the literature is the relative size of the CPRD and length of follow-up. Having reviewed the other studies identified by the systematic review in Chapter **??**, this analysis of 1,684,564 participants is one of the largest available studies of this research question. Additionally, this analysis followed LRA users and non-users from a common index date, using a time-updating treatment indicator to correctly assign time-at-risk to the exposed and unexposed groups. This approach has been less commonly used in the literature and allows for the mitigation of potential immortal time bias.

However, the findings of this analysis are subject to several limitations in addition to the confounding by indication discussed above. There is a strong possibility of differential misclassification[39] of dementia-related conditions based on the exposure.

As an illustrative example, those with memory complaints may be more likely to be classified as vascular dementia than Alzheimer's disease if their medical records contains prescriptions for lipid-regulating agents. Further, there is a potential for general non-differential misclassification of the outcome based on varying positive predictive value of electronic health record code lists to identify dementia cases.[40,41]

Misclassification of outcomes is not the only issue introduced by the use of EHR codes to define outcomes. Comparing and contrasting between different studies is particularly difficult because of the impact that the use of different code list can have on the analysis, as illustrated by the discrepancy between the results when using the codes lists defined for this study and those used by Smeeth *et al.* This is a particular challenge in comparing research across different time-periods and coding systems,

A further limitation stems from uncontrolled confounding due to genetic factors. Number of ApoE $\epsilon 4$ alleles represents the strongest genetic risk factor for Alzheimer's disease, but also substantially increases LDL-c levels,[**bennet2007?**] potentially prompting treatment with a statin or other lipid regulating agent. We were unable to control for ApoE genotype in this analysis as we did not have access to genetic data on participants. As a result, any protective association between LRA use and the Alzheimer's disease outcomes may be masked by residual negative confounding by ApoE.

Finally, as with many studies of dementia, there is a risk of reverse causation in our analysis. Dementia and associated conditions have a long prodromal period, during which preclinical disease could cause indications for the prescription of a lipid-regulating agent.

1.5.5 Enabling easy synthesis of this analysis

In light of my own experiences in attempting to extract information for papers assessing preventative treatments, as documented in Section **??**, the outputs from this analysis are readily available.

The raw data supporting this analysis is not available, as access to the CPRD data is controlled by a data monitoring committee. In the context of access-controlled data, sharing the analysis code represents a way for readers to validate the findings.[42] As such, all code, Read code lists and summary statistics (i.e. the tables presented in this chapter, plus summary tables of effect estimates) are readily available in machine readable formats (i.e. as comma separated values, or CSV, files) from the archived repository for this project (Zotero DOI: **TBC**).

This open approach should enable easy inclusion of this analysis in future evidence synthesis exercises, allowing new work to build on that presented here.

1.5.6 Conclusions

This chapter has provided new evidence on the association of lipid-regulating agents with incidence of all-cause dementia, Alzheimer's disease, vascular dementia, and other dementia. It made use of a large scale electronic health record database, the CPRD, and employed a time-varying Cox proportional hazards model to account for the potential immortal time bias.

It found little evidence for an effect of lipid-regulating agents on probable or possible Alzheimer's disease. However, lipid-regulating agent use was associated with an increased risk of all-cause, vascular and other dementias. In all cases, the estimated associations were driven by those observed in the statin subgroup, which comprised the majority of exposed participants in this cohort.

This chapter attempted to account for important sources of bias and provide a comparison with other available literature, as identified in the systematic review presented in Chapter **??**. However, there is a strong potential for uncontrolled confounding by indication and differential misclassification of the outcome on the basis of exposure, which raises questions about the findings, in particular the unexpected increase in risk of vascular dementia associated with statin use. This is supported by our findings for the negative and positive control outcomes used,

which provide some evidence of uncontrolled vascular confounders that may both prompt LRA prescription and increase risk of vascular dementia. Future research using large scale electronic health records should aim to address these limitations, potentially by using a analytical design that more closely emulates a trial.[38]

Regardless, this analysis has provided an additional source of evidence for the triangulation exercise presented in Chapter **??**. In the following Chapter, the dataset described here is incorporated along with several other datasets as part of an IPD analysis to investigate the effect of blood lipid levels on dementia outcomes directly, rather than via the proxy of lipid-regulating treatment.

References

00 1. Dorn, H. F. Philosophy of Inferences from Retrospective Studies. *American Journal of Public Health and the Nations Health* **43**, 677–683 (1953).

2. Suissa, S. Immortal Time Bias in Pharmacoepidemiology. *American Journal of Epidemiology* **167**, 492–499 (2008).

3. Walker, V. M., Davies, N. M., Jones, T., Kehoe, P. G. & Martin, R. M. Can commonly prescribed drugs be repurposed for the prevention or treatment of Alzheimer's and other neurodegenerative diseases? Protocol for an observational cohort study in the UK Clinical Practice Research Datalink. *BMJ Open* **6**, (2016).

4. von Elm, E. *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Journal of Clinical Epidemiology* **61**, 344–349 (2008).

5. Herrett, E. *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology* **44**, 827–836 (2015).

6. Williams, T., van Staa, T., Puri, S. & Eaton, S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Therapeutic Advances in Drug Safety* **3**, 89–99 (2012).

7. Wood, L. & Coulson, R. Revitalizing the general practice research database: Plans, challenges, and opportunities. *Pharmacoepidemiology and drug safety* **10**, 379–383 (2001).

8. Mathur, R. *et al.* Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *Journal of Public Health (Oxford, England)* **36**, 684–692 (2014).

9. Booth, N. What are the Read Codes? *Health Libraries Review* **11**, 177–182 (1994).

10. Wolf, A. *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *International Journal of Epidemiology* **48**, 1740–1740g (2019).

11. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2017).

12. Walker, V. M., Davies, N. M., Martin, R. M. & Kehoe, P. G. Comparison of Antihypertensive Drug Classes for Dementia Prevention. *Epidemiology* **31**, 852–859 (2020).

13. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of chronic diseases* **40**, 373–383 (1987).

14. Wells, B. J., Chagin, K. M., Nowacki, A. S. & Kattan, M. W. Strategies for handling missing data in electronic health record derived data. *Egems* **1**, (2013).

15. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* **338**, b2393 (2009).

16. Moons, K. G., Donders, R. A., Stijnen, T. & Harrell Jr, F. E. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology* **59**, 1092–1101 (2006).

17. Lévesque, L. E., Hanley, J. A., Kezouh, A. & Suissa, S. Problem of immortal time bias in cohort studies: Example using statins for preventing progression of diabetes. *BMJ* **340**, b5087 (2010).

18. Lamarca, R., Alonso, J., Gomez, G. & Munoz, A. Left-truncated Data With Age as Time Scale: An Alternative for Survival Analysis in the Elderly Population. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **53A**, M337–M343 (1998).

19. Gail, M. H., Graubard, B., Williamson, D. F. & Flegal, K. M. Comments on "Choice of time scale and its effect on significance of predictors in longitudinal studies" by Michael J. Pencina, Martin G. Larson and Ralph B. D'Agostino, *Statistics in Medicine* 2007; **26** :1343-1359. *Statistics in Medicine* **28**, 1315–1317 (2009).

20. Pencina, M. J., Larson, M. G. & D'Agostino, R. B. Choice of time scale and its effect on significance of predictors in longitudinal studies. *Statistics in Medicine* **26**, 1343–1359 (2007).

21. Pigott, T. D. A review of methods for missing data. *Educational research and evaluation* **7**, 353–383 (2001).

22. Hughes, R. A., Heron, J., Sterne, J. A. C. & Tilling, K. Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology* **48**, 1294–1304 (2019).

23. Lipsitch, M., Tchetgen, E. T. & Cohen, T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology (Cambridge, Mass.)* **21**, 383–388 (2010).

24. Selva-O'Callaghan, A. *et al.* Statin-induced myalgia and myositis: An update on pathogenesis and clinical recommendations. *Expert review of clinical immunology* **14**, 215–224 (2018).

25. Collins, R. *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet* **388**, 2532–2561 (2016).

26. Herrett, E. *et al.* Statin treatment and muscle symptoms: Series of randomised, placebo controlled n-of-1 trials. *BMJ* **372**, n135 (2021).

27. Macedo, A. F., Douglas, I., Smeeth, L., Forbes, H. & Ebrahim, S. Statins and the risk of type 2 diabetes mellitus: Cohort study using the UK clinical practice pesearch datalink. *BMC Cardiovascular Disorders* **14**, 85 (2014).

28. Karalis, D. G., Hill, A. N., Clifton, S. & Wild, R. A. The risks of statin use in pregnancy: A systematic review. *Journal of Clinical Lipidology* **10**, 1081–1090 (2016 Sep-Oct).

29. Smeeth, L., Douglas, I., Hall, A. J., Hubbard, R. & Evans, S. Effect of statins on a wide range of health outcomes: A cohort study validated by comparison with randomized trials. *British Journal of Clinical Pharmacology* **67**, 99–109 (2009).

30. Carbonari, D. M. *et al.* Use of demographic and pharmacy data to identify patients included within both the Clinical Practice Research Datalink (CPRD) and The Health Improvement Network (THIN). *Pharmacoepidemiology and Drug Safety* **24**, 999–1003 (2015).

31. Chao, T.-F. *et al.* Statins and the risk of dementia in patients with atrial fibrillation: A nationwide population-based cohort study. *International Journal of Cardiology* **196**, 91–97 (2015).

32. Poly, T. N. *et al.* Association between Use of Statin and Risk of Dementia: A Meta-Analysis of Observational Studies. *Neuroepidemiology* **54**, 214–226 (2020).

33. Williams, D. M., Finan, C., Schmidt, A. F., Burgess, S. & Hingorani, A. D. Lipid lowering and Alzheimer disease risk: A mendelian randomization study. *Annals of Neurology* **87**, 30–39 (2020).

34. Chu, C.-S. *et al.* Use of statins and the risk of dementia and mild cognitive impairment: A systematic review and meta-analysis. *Scientific Reports* **8**, 5804 (2018).

35. Ancelin, M.-L. *et al.* Lipid lowering agents, cognitive decline, and dementia: The three-city study. *Journal of Alzheimer's Disease* **30**, 629–637 (2012).

36. Salas, M., Hotman, A. & Stricker, B. H. Confounding by Indication: An Example of Variation in the Use of Epidemiologic Terminology. *American Journal of Epidemiology* **149**, 981–983 (1999).

37. Suttorp, M. M., Siegerink, B., Jager, K. J., Zoccali, C. & Dekker, F. W. Graphical presentation of confounding in directed acyclic graphs. *Nephrology Dialysis Transplantation* **30**, 1418–1423 (2015).

38. Danaei, G., García Rodríguez, L. A., Cantero, O. F., Logan, R. & Hernán, M. A. Observational data for comparative effectiveness research: An emulation of randomised trials to estimate the effect of statins on primary prevention of coronary heart disease. *Statistical methods in medical research* **22**, 70–96 (2013).

39. Porta, M. *A dictionary of epidemiology.* (Oxford University Press, 2014).

40. Wilkinson, T. *et al.* Identifying dementia cases with routinely collected health data: A systematic review. *Alzheimer's & Dementia* **14**, 1038–1051 (2018).

41. McGuinness, L. A., Warren-Gash, C., Moorhouse, L. R. & Thomas, S. L. The validity of dementia diagnoses in routinely collected electronic health records in the United Kingdom: A systematic review. *Pharmacoepidemiology and Drug Safety* **28**, 244–255 (2019).

42. Goldacre, B., Morton, C. E. & DeVito, N. J. Why researchers should share their analytic code. *BMJ* **367**, l6365 (2019).