BMJ Open

Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: a cohort study with nested case–control analysis

David Reeves,[1,2] David A Springate,[1,2] Darren M Ashcroft,[3] Ronan Ryan,[4] Tim Doran,[5] Richard Morris,[6] Ivan Olier,[1,7] Evangelos Kontopantelis[1,8]

To cite: Reeves D, Springate DA, Ashcroft DM, *et al*. Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: a cohort study with nested case–control analysis. *BMJ Open* 2014;**4**:e004952. doi:10.1136/bmjopen-2014-004952

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2014-004952).

Received 28 January 2014
Revised 12 March 2014
Accepted 13 March 2014

For numbered affiliations see end of article.

Correspondence to
Dr David Reeves;
david.reeves@manchester.ac.uk

ABSTRACT

Objective: To conduct a fully independent and external validation of a research study based on one electronic health record database, using a different electronic database sampling the same population.

Design: Using the Clinical Practice Research Datalink (CPRD), we replicated a published investigation into the effects of statins in patients with ischaemic heart disease (IHD) by a different research team using QResearch. We replicated the original methods and analysed all-cause mortality using: (1) a cohort analysis and (2) a case-control analysis nested within the full cohort.

Setting: Electronic health record databases containing longitudinal patient consultation data from large numbers of general practices distributed throughout the UK.

Participants: CPRD data for 34 925 patients with IHD from 224 general practices, compared to previously published results from QResearch for 13 029 patients from 89 general practices. The study period was from January 1996 to December 2003.

Results: We successfully replicated the methods of the original study very closely. In a cohort analysis, risk of death was lower by 55% for patients on statins, compared with 53% for QResearch (adjusted HR 0.45, 95% CI 0.40 to 0.50; vs 0.47, 95% CI 0.41 to 0.53). In case-control analyses, patients on statins had a 31% lower odds of death, compared with 39% for QResearch (adjusted OR 0.69, 95% CI 0.63 to 0.75; vs OR 0.61, 95% CI 0.52 to 0.72). Results were also close for individual statins.

Conclusions: Database differences in population characteristics and in data definitions, recording, quality and completeness had a minimal impact on key statistical outputs. The results uphold the validity of research using CPRD and QResearch by providing independent evidence that both datasets produce very similar estimates of treatment effect, leading to the same clinical and policy decisions. Together with other non-independent replication studies, there is a nascent body of evidence for wider validity.

Strengths and limitations of this study

- Previous comparisons of electronic health record (EHR) databases have compared different patient populations or have not been done by independent researchers. This is the first fully independent validation of a published EHR-based study using a different EHR database sampling from the same underlying population.
- Estimates obtained from Clinical Practice Research Datalink (CPRD) for the treatment effects of statins on mortality in patients with ischaemic heart disease (IHD) were remarkably similar to those from QResearch, providing a degree of reassurance for clinicians, researchers and policy-makers that findings using either PCD would be essentially the same.
- There were some demographic and other differences between the CPRD and QResearch IHD cohorts. Sensitivity analysis indicated that these had only a minimal effect on the results.
- We were able to successfully replicate nearly all the elements of the original QResearch study using CPRD, but this would not have been possible without some input on methodological detail from the authors of the original study.
- The results add to evidence for the wider validity of the UK primary care databases, but cannot be generalised to EHRs in other countries where the data quality may be quite different.

INTRODUCTION

In recent years, electronic patient health records have emerged as an important new tool for medical research. Numbers of research publications based on analyses of electronic record databases have increased rapidly, and government-sponsored research networks—such as the Observational Medical

Outcomes Partnership (OMOP) in the US and Canadian Network for Observational Drug Effect Studies (CNODES)—have been established to advance research based on electronic records.

Globally, some of the largest and most detailed sources of electronic patient data are the UK 'primary care databases' (PCDs), some of which contain detailed data on all primary care consultations for millions of patients and span more than two decades. The three major UK PCDs are the CPRD[1] (Clinical Practice Research Datalink; formerly the general practice research database (GPRD)), QResearch[2] and The Health Improvement Network[3] (THIN). These PCDs, and CPRD in particular, are used as a resource by researchers throughout the world. A search on PubMed reveals that the combined number of articles based on data from these three PCDs published during 2000 was 41; during 2012, it was 172.

Despite the growing use of electronic datasets as research tools, there remain concerns about the validity of studies based on such data, including uncertainties around data quality, data completeness and the potential for bias from measured and unobserved confounders. The validity of coding in the UK PCDs has received a considerable amount of attention in the literature,[4–7] in particular the completeness of recording of consultations and disease diagnoses. The former has been found to be generally high,[4] whereas the coding of diagnoses varies considerably by condition.[6] The validity of data on risk factors—such as blood pressure and cholesterol—has received less attention,[8–10] but is of particular importance for PCD-based effectiveness studies, where selection bias and non-random missing data have the potential to produce misleading conclusions.[8]

An alternative approach to assessing the validity of PCD-based studies is to compare the results with those obtained from equivalent investigations conducted on other independent datasets. This approach subsumes concerns about the validity and completeness of the underlying individual data values within the broader question of whether such flaws in the data make any difference to the ultimate conclusions drawn from the analysis. A number of studies have compared results obtained from PCDs with those from pre-existing randomised controlled trials (RCTs) of the same intervention.[11–13] However, the use of RCTs as a gold standard for judging the validity of PCD-based studies is questionable: results may differ due to the very different contexts of RCTs and observational data; for example, drug regimes, patient characteristics and use of additional or combination medications may differ radically in the real-world setting. Another approach to validating a PCD-based study is to replicate the study in one or more completely independent PCDs. Agreement of results, although not proving validity, would imply that the conclusions do not depend on the data source. This is the approach we take in the present paper. Several studies have applied the same design protocol to more than one database.[14–19] In most cases, these databases have covered different countries,[15–19] different geographical areas of the same country[16 17] or different patient populations within a country,[14] and also different kinds of databases (eg, administrative claims data vs electronic health records[14]). Some studies have reported consistent findings across databases; others have found varying results. In a study examining the heterogeneity of effect estimates for 53 drug-outcome pairs across 10 US databases (either claims data or electronic health records) using two different study designs,[14] only nine pairs were consistent across databases in direction and statistical significance, with up to 19 pairs having effect estimates ranging from significantly increased risk to significantly decreased risk.

With these studies, it is not possible to determine whether the heterogeneity of results is due to differences in data recording and quality between databases, or to differences in demographics and health between the covered populations. To address this, comparisons are required involving databases that sample from the same underlying patient population. A few studies fall into this category, all based on UK PCDs. Bremner *et al*[20] examined early-life exposure to antibacterials and subsequent development of hay fever by running identical analyses on the CPRD and the Doctors' Independent Network (DIN-LINK). The two child cohorts proved similar in all essentials and results of case-control analysis were similar for both, even extending to a significant association between antibacterial exposure and development of hay fever disappearing after adjustment for a number of consultations. Vinogradova *et al*[21] examined the relationship between exposure to bisphosphonates and gastrointestinal cancers using QResearch and CPRD data. They reported the two patient samples to be similar in demographics, risk factors, comorbidities and use of medications, and found no significant associations between bisphonosphonate use and various types of cancers in either database.

Both of these studies, although informative, were however conducted by research groups instrumental in the development of the comparative PCD (DIN-LINK and QResearch, respectively) and so lacked independence. The only fully independent studies are a series of external validation studies using the THIN database and risk prediction tools originally developed using QResearch (eg, QRISK, QRISK2, QKidney, QStatin).[10 22 23] These studies applied the risk algorithms previously derived using QResearch on patients in THIN, and reported mostly good discriminative and calibration properties. However, these studies did not address the question of whether analysis of the two databases would result in the same at-risk algorithm itself. In this paper, we report what we believe to be the first completely independent full replication of a published research study based on one PCD, using a different PCD that samples from the same population. Our overall aim was to assess the extent to which the model parameters

(in this case, treatment effects) derived using one PCD would be identical to those derived using the other PCD. We also examine a different clinical topic and outcome from these previous replications.

METHODS

CPRD and THIN obtain their data from practices using the Vision electronic record system, while QResearch obtains data from practices using EMIS software. We felt that comparisons would be most informative between databases drawing data from different capture systems. Across the time-period studied, two versions of EMIS were in use, the more common[24] being the text-based EMIS LV system with navigation and data entry mainly via the keyboard; EMIS PCS, which is Windows-based with mouse control and drop-down menus, was introduced from 1999. Vision was Windows-based throughout the study period. A small-scale direct comparison of EMIS LV and Vision indicated that coded data entry, excepting prescribing information, was faster with Vision and that more items were likely to be coded.[25] Practices running Vision have slightly higher achievement rates for most Quality and Outcomes Framework (QOF) indicators than practices running either version of EMIS, even after controlling for differences in practice and area characteristics.[24] We had access to CPRD, and therefore chose to replicate a study previously conducted using QResearch. CPRD and QResearch both draw data from general practices spread throughout the UK—currently more than 600 practices each—and comparisons to the national age-gender structure and prevalence rates for common conditions mostly show good correspondence for both datasets.[26 27] For practical reasons, we focused on studies of the effectiveness of medicinal interventions and, after assessing the available studies, chose to replicate an investigation into the effects of statins on the mortality of patients with ischaemic heart disease (IHD) by Hippisley-Cox and Coupland (H-C&C).[28] The methodological details provided in the published paper were insufficient on their own to allow a close replication to be conducted, and we therefore obtained additional details from the authors. We requested purely factual information about the methods used and did not share any of our analyses or results.

We replicated the methods of H-C&C as closely as possible, given the differences between the two databases. All of the methods described below, including the study period, variable specifications and analytical procedures, are exact replications of those used in the original study, unless indicated otherwise. We selected all practices in CPRD that provided up to standard (UTS) data (UTS is CPRD's designation for data meeting their internal quality standards) for the whole of the period from 1 January 1996 to 17 December 2003. We next identified all patients with a first diagnosis of IHD within this period, based on the QOF business rules for 2004.[29] We excluded patients whose IHD diagnosis fell within the first 3 months of registration with their general practice or was on or subsequent to their recorded date of death, or who were prescribed statins prior to first diagnosis.

We extracted data for these patients from the date of IHD diagnosis up until 17 December 2003, or until the date of death or exit from the practice, or the last recorded date for practices that stopped providing data before 17 December 2003, giving a maximum possible length of follow-up postdiagnosis of just under 8 years.

Analysis

The main outcome was all-cause mortality, identified through a record of death in the CPRD. Following H-C&C, we conducted two main analyses: (1) a cohort analysis and (2) a case-control analysis nested within the full cohort. All analyses were conducted using R.[30] Following H-C&C, statistical significance was assessed using $p<0.01$ (two tailed), but 95% CIs are reported in tables and figures.

We made an a priori decision not to attempt to 'improve' on the analysis conducted by H-C&C, as our specific aim was to determine whether the same results and conclusions would emerge from using identical methods on a different underlying dataset targeting the same population.

Cohort analysis

The cohort analysis used a Cox proportional hazards model to examine the effect of statin use on patient survival, with survival time determined by the time (in days) between the date of first diagnosis and date of death. Patients who transferred out of their practice before death or who were still alive at the end of the study period were treated as censored observations. Statin exposure was used as a time-varying covariate, with the period of exposure from the date of first prescription to when the statin was stopped (estimated as the date of last prescription plus 90 days; intervening breaks in the use of statins were ignored), or if not stopped until the end of the study period, date of death or date of transfer out of practice. Covariates adjusted for in the analysis were year of diagnosis, gender, comorbidities (diabetes, hypertension, myocardial infarction, congestive cardiac failure and cancer), and age (coded as 0–44, 45–54, 55–64, 65–74, 75–84, 85–94 or ≥95), smoking (ever smoked, never smoked, not recorded) and body mass index (BMI; coded as <25, 25–30, >30 kg/m^2) all at the date of diagnosis. The presence of each comorbidity was indicated by a diagnosis in the patient record (using the 2004 QOF business rules) and coded as present/not present at the date of IHD diagnosis. If smoking status or BMI was not recorded within 4 years prior to diagnosis of IHD, we coded it as missing.

The analysis was undertaken using the R survival analysis package accounting for the clustering of patients by practice and using the Huber-White robust estimate of SE. The proportional hazards assumption was checked graphically and with a test for proportional hazards.

Nested case-control study

The nested case-control analysis compared all patients from the cohort who died during the follow-up period (the cases) with a group of matched control patients (also with IHD) who did not die. For each case, we defined an 'index date' as the date of death. We then used an incidence density sampling procedure (as per the original study; personal correspondence) to randomly select four control patients for each case matched on gender, year of IHD diagnosis and age (coded in 5-year age-bands). General practice was not used as a matching variable. Controls were patients with IHD alive at the time their matched case died (including patients who themselves became cases at a later time-point). The incidence sampling procedure allowed the same patient to be selected as a control for more than one case, thus providing a full set of four controls for each case, while still producing unbiased estimates of risk.[31]

Statin exposure was based on the first and last prescription dates prior to the index date and coded into: (1) currently taking statins (last prescription was within 90 days of the index date); (2) previously took statins (last prescription more than 90 days prior to the index date) and (3) has never taken statins. We did this for all statins as a group and also separately for five different types of statin (atorvastatin, cerivastatin, fluvastatin, pravastatin and simvastatin). For 'all statins', the last prescription could be for a different statin type than the first; for individual statins, it had to be the same type. One further formulation, rosuvastatin, was in use that did not appear in the QResearch study. We included this in the 'all statins' group but did not analyse it individually as only 22 patients had received the statin.

Analysis of the case-control study used conditional logistic regression accounting for the matching of cases with controls, to obtain ORs for the risk of death in relation to use of statins. We allowed for clustering by general practice and used a robust estimate of SE, in line with the cohort analysis. Covariates in the analysis were smoking status, BMI and comorbidities, specified as in the Cohort analysis but based on the index date rather than the date of diagnosis. Additional covariates in this analysis were the Townsend deprivation score for the practice postcode (in national quintiles; H-C&C used quintiles of patient-level Townsend scores) and use of β-blockers, aspirin, ACE inhibitors and calcium channel blockers, identified through the British National Formulary[32] chapter codes in the patient record. Each medication was coded as either used or not used prior to the index date but after the date of IHD diagnosis. Interactions between use of statins and each of gender, age (less than 75 vs 75 and over) and diabetes were tested by adding interaction terms into the model.

Sensitivity analysis

To replicate the original sensitivity analyses, we reran the case-control study: (1) while excluding patients without recorded values for BMI or smoking; (2) with the sample restricted to patients without a diagnosis of cancer; (3) with the sample restricted to patients without diabetes, congestive cardiac failure or myocardial infarction and (4) using only those cases who survived for at least 1 year after diagnosis of IHD and their matched controls. The definitions of death and deprivation were different in CPRD and to assess sensitivity to this we repeated the cohort and case-control analyses with the analyses restricted to practices for which patient-level Office of National Statistics (ONS) official death dates and Townsend scores were available (58% of practices and 60% of patients).

Our primary analysis replicated H-C&C in restricting the sample to only those practices with data for the full 8-year period. However, inclusion criteria for CPRD-based studies are generally patient-based rather than practice-based, and include all individual patients with UTS data for the analysis period (ie, from diagnosis date to end of study, death or transfer out of practice), and on this basis 61 458 patients from 577 CPRD practices could be included. We therefore repeated the main analyses using this sample.

RESULTS

Comparison of patient cohorts

A higher number of practices contributed data to the CPRD cohort: 224 compared with 89 for QResearch, resulting in a total sample of patients with a first diagnosis of IHD in the study period, after exclusions, of 34 925 compared with 13 029 (table 1; note that if the original study were undertaken now, additional practices (with their retrospective data) added to QResearch since 2006 would produce a more equivalently sized cohort). Incidence cases per practice were considerably higher for CPRD (on average, 242 compared with 190), possibly implying that the included CPRD practices were generally larger, though a smaller proportion of CPRD patients met the study inclusion criteria (64.4% vs 77%). H-C&C provided descriptive statistics for only certain covariates, and reported these in person-years rather than counts. Total person-years of observations in CPRD were 125 709 compared with 43 460 in QResearch. QResearch included a greater proportion of person-years from older patients (36.3% from patients aged 75 or over compared with 28.1%) and had a much higher representation of congestive cardiac failure (14% compared with 6.0%), but less hypertension (28.9% compared with 35.9%). These figures imply some demographic differences between the cohorts.

Table 2 reports mortality rates from the two studies by various patient characteristics. Age-band specific mortality rates were higher in CPRD for all age-bands except the youngest (0–44 years), although, owing to differing age distributions, the overall mortality rates were very similar (53.5/1000 person-years for CPRD compared with 52.1 for QResearch). Mortality was slightly higher

Table 1 Descriptive statistics of the CPRD and QResearch samples

	CPRD		QResearch	
Number of general practices in the study	224		89	
All patients with a first diagnosis of IHD during the study period (incidence rate per practice)	54 217 (242)		16 920 (190)	
Number (%) of incident cases meeting the inclusion criteria	34 925 (64.4%)		13 029 (77.0%)	
End of study status				
Died	6725 (19.3%)		2266 (17.4%)	
Alive	24 292 (69.6%)		9609 (73.8%)	
Left before end of study	3908 (11.2%)		1154 (8.9%)	
Total person years of observation	125 709		43 460	
	Person years	%	Person years	%
Age band (years)				
0–44	4997	4.0	824	1.9
45–54	16 506	13.1	3923	9.0
55–64	30 431	24.2	9270	21.3
65–74	38 365	30.5	13 636	31.4
75–84	28 082	22.3	11 827	27.2
85–94	7073	5.6	3744	8.6
≥95	255	0.2	235	0.5
Women	57 169	45.5	18 539	42.7
Men	68 540	54.5	24 920	57.3
No diabetes	114 949	91.4	39 814	91.6
Diabetes	10 760	8.6	3646	8.4
No hypertension	80 574	64.1	30 912	71.1
Hypertension	45 136	35.9	12 547	28.9
No congestive cardiac failure	118 209	94.0	37 391	86.0
Congestive cardiac failure	7501	6.0	6069	14.0

CPRD, Clinical Practice Research Datalink; IHD, ischaemic heart disease.

for women than for men in both datasets. In both cohorts, diabetes and congestive cardiac failure were associated with greatly increased death rates.

Survival analysis

The Kaplan-Meier survival curve, uncontrolled for any covariates (figure 1), shows a clear survival advantage

Table 2 Comparison of mortality rates in the Clinical Practice Research Datalink (CPRD) and QResearch cohorts: eligible patients only

	CPRD				QResearch	
Cohort	Person years	Number of deaths	Rate/1000 person-years	95% CI	Rate/1000 person-years	95% CI
All	125 709	6725	53.5	52.3 to 54.8	52.1	50 to 54.3
Age band (years)						
0–44	4997	44	8.8	6.5 to 11.9	9.7	4.9 to 19.4
45–54	16 506	202	12.2	10.6 to 14.1	10.2	7.5 to 13.9
55–64	30 431	638	21.0	19.4 to 22.7	16.8	14.4 to 19.7
65–74	38 365	1639	42.7	40.7 to 44.8	32.8	29.9 to 36
75–84	28 082	2638	93.9	90.6 to 97.4	77.0	72.2 to 82.2
85–94	7073	1457	206.0	196.6 to 215.6	167.2	154.6 to 180.8
≥95	255	107	419.8	359.0 to 483.1	331.4	265.4 to 413.7
Women	57 169	3116	54.5	52.7 to 56.4	54.1	50.9 to 57.6
Men	68 540	3609	52.7	51.0 to 54.4	50.7	48 to 53.6
No diabetes	114 949	5835	50.8	49.5 to 52.1	49.7	47.5 to 51.9
Diabetes	10 760	890	82.7	77.6 to 88.1	79.0	70.4 to 88.7
No hypertension	80 574	4241	52.6	51.1 to 54.2	50.8	48.3 to 53.4
Hypertension	45 136	2484	55.0	53.0 to 57.2	55.5	51.5 to 59.8
No CCF	118 209	5410	45.8	44.6 to 47.0	41.4	39.3 to 43.5
CCF	7501	1315	175.3	166.8 to 184.2	118.6	110.3 to 127.6

CCF, congestive cardiac failure.

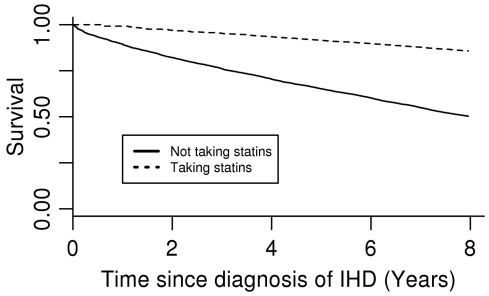

Figure 1 Kaplan-Meier plot showing survival of patients taking statins compared with patients not taking statins, uncontrolled for covariates; CPRD study.

for patients taking statins, with a raw HR of 0.25 (95% CI 0.23 to 0.27; table 3). At 6 years, the unadjusted survival rate for patients taking statins was 89%, versus 63% for those not taking statins, remarkably similar to the reported values of 89% and 66% for the QResearch cohort.

The Cox proportional hazards model (adjusted for the covariates gender, age, year of diagnosis, diabetes, hypertension, congestive cardiac failure, myocardial infarction, cancer, BMI and smoking) significantly departed from the proportional hazards assumption when year of diagnosis was specified as a continuous variable (p<0.01). Respecifying year of diagnosis as a stratification variable resolved this problem. The adjusted HR was 0.45 (95% CI 0.40 to 0.50), very close to the adjusted HR for QResearch patients of 0.47 (95% CI 0.41 to 0.53) and representing a 55% lower risk of death for patients on statins.

Case-control study

The case-control analysis included 6683 cases and 26 732 controls (for 42 cases, we were unable to find matching controls). The cases and controls were well matched in terms of median age, gender and duration of IHD (table 4). They were also very close in these respects to the patients in the QResearch study. Compared with QResearch, slightly lower percentages of cases and controls had received a prescription for statins (16.6% vs 19.6% and 22.8% vs 25.4%, respectively), though there were smaller differences in the percentages taking them for more than 1 year. For completeness, online supplementary table A provides a comparison of the CPRD cases and controls on the unmatched covariates in the analysis (not available for QResearch). These show a good to acceptable balance on all covariates.

Table 3 Unadjusted and adjusted HRs (95% CI) of risk of death for patients taking statins compared with patients not taking statins; Clinical Practice Research Datalink (CPRD) and QResearch studies

	CPRD	QResearch
Unadjusted	0.25 (0.23 to 0.27)	Not reported
Adjusted	0.45 (0.40 to 0.50)	0.47 (0.41 to 0.53)

Patients who were currently on a statin had a significantly decreased rate of death compared with patients who had never taken a statin in unadjusted (OR=0.57, 95% CI 0.53 to 0.62) and adjusted (OR=0.69, 95% CI 0.63 to 0.75) analyses (table 5). These ORs were very close to those reported for QResearch (unadjusted OR=0.53, 95% CI 0.46 to 0.61; adjusted OR=0.61, 95% CI 0.52 to 0.72; figure 2). Patients who were previously, but not currently, on a statin had a significantly elevated risk of death in unadjusted and adjusted analysis, whereas H-C&C found a significantly increased risk in unadjusted analysis only.

In the case of individual statins, the QResearch study reported significant protective effects against risk of death for current use of atorvastatin and simvastatin; with CPRD we found the same, but also a protective effect for pravastatin, which just failed to reach 1% significance in QResearch (p=0.013). For all five statins, our current use OR point estimates were similar to those from QResearch, though the larger CPRD sample produced narrower CIs. Like the original study, we found no significant effects in patients who were previously, but not currently, on any individual statin, despite the increased risk for all statins combined.

Effects of age, sex and diabetes on the effectiveness of statins

Like H-C&C, we found no evidence for an interaction between gender and statin use (p=0.84), or diabetes and statin use (p=0.62). Unlike H-C&C, however, we did find a significant interaction with age (p<0.001), with an adjusted OR of 0.55 (95% CI 0.48 to 0.62) for people aged less than 75 and 0.77 (95% CI 0.65 to 0.92) for those aged 75 or over, indicating greater benefit for those under 75 years of age.

Results for sensitivity analyses

Repeating the sensitivity analyses from the original study, we found that restricting the case-control sample to patients with BMI and smoking status information, or to those without cancer, or without a diagnosis of diabetes, congestive cardiac failure or myocardial infarction made very little difference to our results. Restricting the sample to patients alive for at least 1 year after diagnosis of IHD likewise made little difference (see online supplementary figure A).

Restricting the CPRD sample to practices for which patient-level Townsend scores and ONS official death dates were available made no appreciable difference to the results of either the cohort (HR 0.47, 95% CI 0.40 to 0.54) or case-control analyses (see online supplementary figure B). Similarly, widening the sample to include all patients with UTS data made little difference (cohort study HR=0.43, 95%CI=0.39 to 0.47; for Case-Control analysis, see online supplementary figure C).

Table 4 Comparison of cases and controls for CPRD and QResearch studies

	CPRD	QResearch
Number of patients		
Cases	6683	2266
Controls	26 732	9064
Median age in years		
Cases	80	80
Controls	80	80
Male (%)		
Cases	53.6	55.7
Controls	53.6	55.7
Median duration of IHD (months)		
Cases	21.3	20.3
Controls	21.4	21.0
N (%) prescribed statins		
Cases	1108 (16.6)	445 (19.6)
Controls	6083 (22.8)	2303 (25.4)
Of those prescribed statins, N (%) taking them for >12 months		
Cases	572 (51.6)	228 (51.2)
Controls	3398 (55.9)	1336 (58)

CPRD, Clinical Practice Research Datalink; IHD, ischaemic heart disease.

DISCUSSION

We conducted an independent replication of a primary care database-based study using a different primary care database, sampling from the same population. We replicated the methods of the original study as closely as we could and reached exactly the same clinical conclusions concerning the effects of statins on mortality in all their essentials. Not only that, our point estimates for the key statistical parameters—the HR from the cohort analysis, and the ORs from the nested case-control study—were remarkably similar to those reported by the original study. For the period under study, CPRD provided a much larger sample than was included in the original QResearch study.

LIMITATIONS

While we were able to exactly replicate nearly all the elements of the original study, there were a few minor differences due mainly to data specifications. The datasets may have differed in the way in which all-cause mortality is defined, as both use their own bespoke algorithm. For area deprivation, QResearch used Townsend quintile deprivation scores at the patient-level, whereas these scores were fully available in CPRD only at the practice level, and for only 60% of the cohort at the patient-level. We tested for the impact of these factors by running a sensitivity analysis using the subset of CPRD patients for which linked ONS data on the date of death and residential Townsend scores were available.

FINDINGS

We observed a number of differences between the CPRD and QResearch cohorts, in particular eligibility rates, age distributions and some comorbidities; hence, there may have been some population differences between the cohorts. Although both PCDs purport to provide nationally representative coverage of the UK population, only subsets of the practices in each were included (those providing data for the whole study period) and it may be that these differed in coverage: for the time-period of the study QResearch included more practices from the South and East of the UK[33] than our CPRD cohort, whereas the latter included higher concentrations of practices from London, the North West and Scotland. It is also possible that recording differences between the Vision and EMIS computer software systems[24] resulted in differential coding of some comorbidities.

Despite these differences, all of our key results using CPRD were very close to those based on QRresearch. The cohort analysis returned remarkably similar values

Table 5 Unadjusted and adjusted ORs comparing cases and controls by type and use of statins, CPRD study

	Unadjusted OR (compared with never used)	95% CI	Adjusted OR (compared with never used)	95% CI	p Value
Used previously					
Any statin	1.43	1.24 to 1.65	1.4	1.21 to 1.63	<0.001
Atorvastatin	1.15	0.92 to 1.45	1.18	0.94 to 0.50	0.158
Cerivastatin	0.59	0.43 to 0.83	0.67	0.48 to 0.94	0.02
Fluvastatin	0.77	0.50 to 1.19	0.89	0.56 to 1.40	0.611
Pravastatin	1.03	0.78 to 1.36	1.13	0.85 to 1.50	0.398
Simvastatin	1.02	0.87 to 1.19	1.07	0.91 to 1.26	0.385
Used currently					
Any statin	0.57	0.53 to 0.62	0.69	0.63 to 0.75	<0.001
Atorvastatin	0.55	0.48 to 0.64	0.66	0.57 to 0.77	<0.001
Cerivastatin	0.48	0.30 to 0.77	0.59	0.36 to 0.96	0.0336
Fluvastatin	0.50	0.33 to 0.76	0.65	0.43 to 0.99	0.0468
Pravastatin	0.60	0.49 to 0.73	0.68	0.56 to 0.84	<0.001
Simvastatin	0.64	0.58 to 0.71	0.78	0.70 to 0.87	<0.001

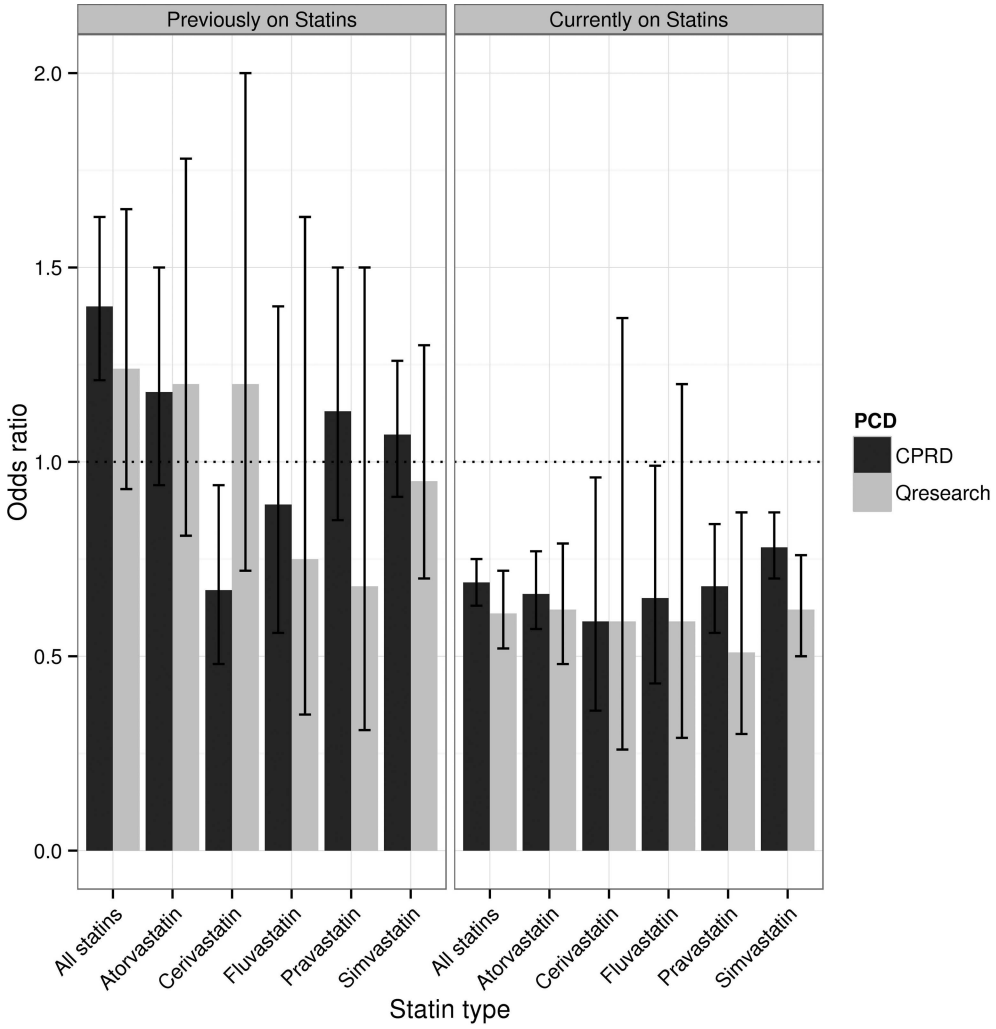

Figure 2 Adjusted ORs comparing cases and controls by type and use of statin; CPRD and QResearch studies.

to H-C&C for the statin use HR, both with and without control for covariates. Likewise, the nested case-control study yielded a very similar estimate for the protective effect of current statin use. Also, like H-C&C, we found separate protective effects for current users of atorvastatin and simvastatin. Another formulation, pravastatin, was found to be protective using CPRD but just failed to reach 1% significance in the QResearch study, the difference most likely being due to the larger CPRD sample.

In both studies, evidence for an elevated risk of death for patients who were previously, but not currently, on statins was somewhat at odds with the results for individual statins. However, cerivastatin was withdrawn from the market in 2001, in the middle of the study period,[34] which may have had a complex impact on these results, particularly since cerivastatin users are likely to have switched to a different statin on removal. To examine the impact of discontinuation of cerivastatin, we repeated the case-control analysis using only the data prior to 1 January 2001. The resulting all-statins OR was no longer statistically significant and in greater accord with those for individual statins (see online supplementary figure D).

Findings from meta-analyses of a large number of RCTs leave little doubt that statins do indeed benefit patients with IHD.[35 36] However, the more than 50% reductions in mortality risk in CPRD and QResearch are very much greater than the reductions of 20–30% reported in major trials,[37–39] or the overall reduction of 17% from meta-analysis of 92 trials.[34] One possibility is that these observational studies are biased by unmeasured confounding factors, but another is that RCTs might substantially underestimate the benefit of statins in the actual population of users.

However, our intention in conducting the research was more methodological than clinical: to establish whether analyses of different PCDs would lead to the same overall clinical conclusions. To this end, we kept all aspects of the analysis as constant as possible except for the PCD itself. The closeness of the results suggests that any variations between the datasets in population characteristics, data definitions, data quality and completeness had only a very minimal impact on the key statistical outputs: the HRs and ORs that are the main parameters used to inform clinical and policy decision-making. The few differences in statistical significance

were principally attributable to the considerably larger size of the CPRD sample.

Our results also demonstrate that PCD-based studies can be successfully and independently replicated in other PCDs. However, this was only possible with the cooperation of H-C&C, as the original paper did not include the necessary methodological detail: for example, the Read and other codes used to define IHD and other morbidities; how drug exposures were measured; the precise specification of each covariate; and the method used to select matched controls. Such absence of methodological detail is near ubiquitous throughout the field and at least partly attributable to journal restrictions on paper length. Most scientific journals now allow supplementary material to be published online alongside the main paper, and we would encourage researchers to publish their full methods online, whenever possible, and journal editors to encourage this. To facilitate this, we have setup an online code-list repository.[40]

Our results provide a degree of reassurance about the validity of PCD-based studies, at least in terms of research undertaken using CPRD and/or QResearch. Together with Vinogradova *et al*'s replication[21] of a different clinical topic, the findings suggest that these two PCDs produce estimates of treatment effects that are substantially the same. Combined with replication studies comparing CPRD and DIN-LINK,[20] there is a nascent body of evidence for wider validity. We also note that whereas previous replications concerned null (non-significant) findings, the present study is evidence for successful replication of a positive intervention effect, which is arguably a stronger test of agreement. However, we emphasise that this paper has addressed validity only in the sense of consistency of statistical results, not the accuracy of the effect estimates relative to some 'true' value or the validity of the clinical conclusions drawn from these: analyses from both PCDs could conceivably be biased in the same direction, due to unmeasured factors common to both or limitations in the analysis methods themselves.[41]

Nevertheless, further replication studies similar to ours are needed. PCDs are used to address a wide range of different kinds of research questions, using a great variety of designs and analytical methods, and replications of studies based around other forms of research design would be particularly informative. Our study used UK PCDs, which are generally acknowledged to be of higher quality and completeness than databases available for most countries, and we would urge researchers in other countries to undertake similar comparison studies.

Author affiliations
[1]NIHR School for Primary Care Research, Centre for Primary Care, Institute of Population Health, University of Manchester, Manchester, UK
[2]Centre for Biostatistics, Institute of Population Health, University of Manchester, Manchester, UK
[3]Centre for Pharmacoepidemiology and Drug Safety Research, Manchester Pharmacy School, University of Manchester, Manchester, UK
[4]Primary Care Clinical Sciences, School of Health and Population Sciences, University of Birmingham, Birmingham, UK
[5]Department of Health Sciences, University of York, York, UK
[6]Department of Primary Care and Population Health, Institute of Epidemiology and Health, University College London, London, UK
[7]Institute of Biotechnology, School of Computer Science, University of Manchester, Manchester, UK
[8]Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester, UK

Acknowledgements This study is based on data from the Clinical Practice Research Datalink (CPRD) obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. However, the interpretation and conclusions contained in this paper are those of the authors alone.

Contributors DR, EK, RR and RM developed the original idea for the study and DS, DA and TD contributed to the study design. DR supervised all aspects of the study's execution. DS helped to plan the analysis and undertook the primary analysis. DR and DS wrote the first draft of the paper. All authors critically reviewed the paper and approved the submitted version. DR is the guarantor of this work and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding This work was funded by the National Institute for Health Research School for Primary Care Research (project number 142).

Competing interests DR was partly funded and DS and IO fully funded by an NIHR School for Primary Care Research grant to undertake this study and EK was partly supported by an NIHR School for Primary Care Research fellowship in primary healthcare.

Ethics approval The study protocol was approved by the independent scientific advisory committee (ISAC) for CPRD research (reference number: 12_149R).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Clinical Practice Research Datalink data cannot be directly shared due to licensing restrictions. All the code-lists used in the analysis of CPRD in this study are available at https://clinicalcodes.rss.mhs.man.ac.uk/. The full R code used for the analysis of CPRD is available from the authors.

REFERENCES

1. Clinical Practice Research Datalink (CPRD). http://www.cprd.com/home/ (accessed Mar 2014).
2. QResearch. http://www.qresearch.org/ (accessed Mar 2014).
3. The Health Improvement Network (THIN) database. http://www.epic-uk.org/ (accessed Mar 2014).
4. Herrett E, Thomas SL, Schoonen WM, *et al*. Validation and validity of diagnoses in the general practice research database: a systematic review. *Br J Clin Pharmacol* 2010;69:4–14.
5. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research database: a systematic review. *BJGP* 2010:e128–36.
6. Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: a systematic review. *Family Pract* 2004;21:396–412.
7. 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.
8. Marston L, Carpenter JR, Walters KR, *et al*. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf* 2010;19:618–26.
9. Delaney JA, Daskalopoulou SS, Brophy JM, *et al*. Lifestyle variables and the risk of myocardial infarction in the general practice research database. *BMC Cardiovasc Disord* 2007;7:38.

10. Collins SC, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010;340:c2442.
11. 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. *Br J Clin Pharmacol* 2009;67:99–109.
12. Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *BMJ* 2009;338:b81.
13. Tannen R, Xie D, Wang X, *et al*. A new 'Comparative Effectiveness' assessment strategy using the THIN database: Comparison of the cardiac complications of pioglitazone and rosiglitazone. *Pharmacoepidemiol Drug Saf* 2013;22:86–97.
14. Madigan D, Ryan PB, Schuemie M, *et al*. Evaluating the Impact of Database Heterogeneity on Observational Study Results. *Am J Epidemiol* 2013;178:645–51.
15. Wijnans L, Lecomte C, de Vries C, *et al*. The incidence of narcolepsy in Europe: Before, during, and after the influenza A (H1N1)pdm09 pandemic and vaccination campaigns. *Vaccine* 2013;31:1246–54.
16. Dormuth CR, Hemmelgarn BR, Paterson JM, *et al*. Use of high potency statins and rates of admission for acute kidney injury: multicenter, retrospective observational analysis of administrative databases. *BMJ* 2013;346:f880.
17. Filion KB, Chateau D, Targownik LE, *et al*. Proton pump inhibitors and the risk of hospitalisation for community-acquired pneumonia: replicated cohort studies with meta-analysis. *Gut* 2014;63:552–8.
18. Valkhoff VE, van Soest EM, Mazzaglia G, *et al*. Adherence to gastroprotection during cyclooxygenase 2 inhibitor treatment and the risk of upper gastrointestinal tract events: a population-based study. *Arthritis Rheum* 2012;64:2792–802.
19. Andrews N, Stoweb J, Miller E, *et al*. A collaborative approach to investigating the risk of thrombocytopenic purpura after measles–mumps–rubella vaccination in England and Denmark. *Vaccine* 2012;30:3042–6.
20. Bremner SA, Carey IM, DeWilde S, *et al*. Early-life exposure to antibacterials and the subsequent development of hayfever in childhood in the UK: case-control studies using the General Practice Research Database and the Doctors' Independent Network. *Clin Exp Allergy* 2003;33:1518–25.
21. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to bisphosphonates and risk of gastrointestinal cancers: series of nested case-control studies with QResearch and CPRD data. *BMJ* 2013;346:f114.
22. Collins SC, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009;339:b2584.
23. Collins SC, Altman DG. Predicting the adverse risk of statin treatment: an independent and external validation of Qstatin risk scores in the UK. *Heart* 2012;98:1091–7.
24. Kontopantelis E, Buchan I, Reeves D, *et al*. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. *BMJ Open* 2013;3:e003190.
25. Refsum C, Kumarapeli P, Gunaratne A, *et al*. Measuring the impact of different brands of computer systems on the clinical consultation: a pilot study. *Inform Prim Care* 2008;16:119–27.
26. Carey IM, Cook DG, De Wilde S, *et al*. Developing a large electronic primary care database (Doctors' Independent Network) for research. *Int J Med Inform* 2004;73:443–53.
27. Hippisley-Cox J, Vinogradova Y. Trends in Consultation Rates in General Practice 1995/1996 to 2008/2009: Analysis of the QResearch database. Final Report to the NHS Information Centre and Department of Health 2009:8. http://www.hscic.gov.uk/catalogue/PUB01077/tren-cons-rate-gene-prac-95–09-95-09-rep.pdf (accessed Mar 2014).
28. Hippisley-Cox J, Coupland C. Effect of statins on the mortality of patients with ischaemic heart disease: population based cohort study with nested case-control analysis. *Heart* 2006;92:752–8.
29. http://www.nice.org.uk/aboutnice/qof/ (accessed Mar 2014).
30. http://www.r-project.org/ (accessed Mar 2014).
31. Richardson DB. An incidence density sampling program for nested case-control analyses. *Occup Environ Med* 2004;61:e59.
32. British National Formulary. http://www.bnf.org/bnf/index.htm (accessed Mar 2014).
33. Hippisley-Cox J, Pringle M, Ryan R. *Stroke: prevalence, incidence and care in general practices 2002 to 2004*. Final report from the QResearch team to National Stroke Audit, 2004. http://www.qresearch.org/Public_Documents/National_Stroke_Audit_Final_Report_2004.pdf (accessed Mar 2014).
34. Furberg CD, Pitt B. Withdrawal of cerivastatin from the world market. *Curr Control Trials Cardiovasc Med* 2001;2:205–7.
35. Cholesterol Treatment Trialists' Collaborators. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. *Lancet* 2005;366:1267–78.
36. Naci H, Brugts JJ, Fleurence R, *et al*. Comparative Benefits of Statins in the Primary and Secondary Prevention of Major Coronary Events and All-Cause Mortality: A Network Meta-Analysis of Placebo-Controlled and Active-Comparator Trials. Supplementary material table 4. *Eur J Prev Cardio* 2013;20:641–57.
37. Scandinavian Simvastatin Survival Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian simvastatin survival study (4S). *Lancet* 1994;344:1383–9.
38. The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. *N Engl J Med* 1998;339:1349–57.
39. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002;360:7–22.
40. https://clinicalcodes.rss.mhs.man.ac.uk/ (accessed Mar 2014).
41. Freemantle N, Marston L, Walters K, *et al*. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;347:f6409.