BMJ Open

Cancer recording in patients with and without type 2 diabetes in the Clinical Practice Research Datalink primary care data and linked hospital admission data: a cohort study

Rachael Williams,[1,2] Tjeerd-Pieter van Staa,[2] Arlene M Gallagher,[1,2] Tarek Hammad,[3,4] Hubert G M Leufkens,[2] Frank de Vries[2]

To cite: Williams R, van Staa T-P, Gallagher AM, *et al*. Cancer recording in patients with and without type 2 diabetes in the Clinical Practice Research Datalink primary care data and linked hospital admission data: a cohort study. *BMJ Open* 2018;**8**:e020827. doi:10.1136/bmjopen-2017-020827

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-020827).

Received 3 December 2017
Revised 31 March 2018
Accepted 3 May 2018

[1]CPRD, Medicines and Healthcare Products Regulatory Agency, London, UK
[2]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
[3]Office of Surveillance and Epidemiology, Food and Drug Administration, Silver Spring, Maryland, USA
[4]EMD Serono Research & Development, EMD Serono, Inc, Rockland, Maine, USA

Correspondence to
Rachael Williams;
Rachael.Williams@mhra.gov.uk

ABSTRACT

Objectives and setting Conflicting results from studies using electronic health records to evaluate the associations between type 2 diabetes and cancer fuel concerns regarding potential biases. This study aimed to describe completeness of cancer recording in UK primary care data linked to hospital admissions records.

Design Patients aged 40+ years with insulin or oral antidiabetic prescriptions in Clinical Practice Research Datalink (CPRD) primary care without type 1 diabetes were matched by age, sex and general practitioner practice to non-diabetics. Those eligible for linkage to Hospital Episode Statistics Admitted Patient Care (HES APC), and with follow-up during April 1997–December 2006 were included.

Primary and secondary outcome measures Cancer recording and date of first record of cancer were compared. Characteristics of patients with cancer most likely to have the diagnosis recorded only in a single data source were assessed. Relative rates of cancer estimated from the two datasets were compared.

Participants 53 585 patients with type 2 diabetes matched to 47 435 patients without diabetes were included.

Results Of all cancers (excluding non-melanoma skin cancer) recorded in CPRD, 83% were recorded in HES APC. 94% of cases in HES APC were recorded in CPRD. Concordance was lower when restricted to same-site cancer records, and was negatively associated with increasing age. Relative rates for cancer were similar in both datasets.

Conclusions Good concordance in cancer recording was found between CPRD and HES APC among type 2 diabetics and matched controls. Linked data may reduce misclassification and increase case ascertainment when analysis focuses on site-specific cancers.

Strengths and limitations of this study

- This study uses a large cohort of patients sourced from the most validated UK primary care database linked to national hospital admissions data.
- The study evaluates recording of cancer across all tumour sites.
- As different coding systems are used in the two data sources, non-concordance may be attributed in part to the challenges in mapping different coding dictionaries.
- The study period was limited by the coverage period of the linked cancer registry data available at the time of the study.

INTRODUCTION

Over 400 million adults have diabetes worldwide, with current estimates suggesting 1 in 10 will live with the disease by 2040.[1] A large number of observational studies that used routinely collected electronic health records (EHRs) have evaluated the association between type 2 diabetes and various types of cancer. However, conflicting results have fuelled concerns regarding the potential for biased associations, including the misclassification of cancer outcomes.[2]

EHRs are increasingly used for observational studies of disease epidemiology and drug safety. The ability to accurately identify cancer events within EHRs would allow for a more valid evaluation of the relative incidences and risks of cancer outcomes in patients with type 2 diabetes, including those exposed to specific antidiabetic medications.[3] However, previous studies of the sensitivity, positive predictive value and agreement between different EHRs for the identification of cancer have demonstrated mixed results.[4–11] Primary care, hospital admissions and disease registry EHRs have each been shown to miss a large proportion of events in other conditions such as myocardial infarction.[12]

Using linked data sources for case ascertainment has been proposed in order to reduce the misclassification of outcomes.[12] Previous research has demonstrated reasonably high concordance between the recording of cancer diagnoses in UK primary care and linked cancer registry data,[13 14] in contrast to results from other countries.[15] Agreement has been shown to vary by cancer site and patient age, meaning misclassification is reduced when linked cancer registration data are used. However, the release of UK cancer registry data for research purposes is subject to time lags due to the current process of validating all the expected registrations for a given calendar year and the associated treatment and outcome information from the following 12 months prior to release.[16] Cancer registrations are almost exclusively based on information supplied by hospitals and from death certification.[13] The objective of this study was to describe the completeness of case ascertainment in the Clinical Practice Research Datalink (CPRD) primary care data linked to Hospital Episode Statistics Admitted Patient Care (HES APC) records, available more contemporaneously than linked cancer registry data. Therefore, the aims of this study were to compare the completeness of recording of cancer, date of first record, characteristics of cases most likely to be missed and relative rates (RRs) of cancer for patients with type 2 diabetes compared with patients without diabetes, across the two datasets.

METHODS

Data sources

The data used for this study were sourced from CPRD, primarily from two routinely collected linked EHR datasets. .

CPRD primary care data comprise the anonymous longitudinal EHR of over 14 million patients from consenting general practitioner (GP) practices in the UK,[17 18] and have been shown in numerous validation studies to be generally of high quality.[19 20] Primary care practitioners are responsible for the management of chronic conditions including type 2 diabetes, and referrals on to specialist care, including for investigation of suspected cancer. Data contain diagnoses made in primary care and records of specialist and secondary care that have been fed back to the GP for the clinical management of the patient, coded using Read diagnosis codes. Free-text notes recorded by GPs or created from scans of letters from specialists were available to access, following anonymisation by CPRD, at the time of the study.

HES APC data include admission and discharge details of all inpatient and day-case admissions in England and Wales from 1997 onwards.[21] HES APC data include all diagnoses for each episode of care within a hospitalisation. The data are validated and cleaned by National Health Service (NHS) Digital at various stages in the processing cycle before derived fields are added and the data made available for research.[22]

In addition, this study used data from official death certificate records sourced from the Office for National Statistics, and cancer registration data sourced from the National Cancer Data Repository.

For the purposes of the current study, the source population was restricted to patients registered with GP practices participating in the CPRD linkage scheme (approximately 60%). CPRD primary care data are routinely linked to other data sources (including HES, death certificates and cancer registration data) at the patient level by NHS Digital, the trusted third party of the CPRD linkage scheme, using patient identifiers stripped from the clinical records. Records from the different data sources are deterministically linked on the basis of the unique patients NHS number, name, gender and postal code of residence. Anonymised linked data are made available to CPRD for the purposes of research, but are not provided back to the GP practices.

Study population

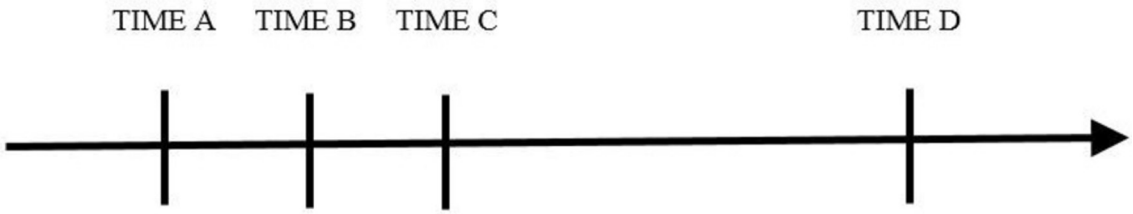
Adult patients aged 40 years and older with type 2 diabetes were identified from primary care records on the basis of one or more prescriptions for insulin or oral antidiabetic medication at least 1 year after the maximum of the patient's registration date with the GP practice and the CPRD derived start date for practice data quality (Up To Standard [UTS] date).[23] The first eligible prescription date was taken as the index date. Patients with a record of type 1 diabetes before the index date were excluded.

Each patient with type 2 diabetes was randomly matched by year of birth (within 5 years), gender and GP practice up to one patient with no records of prescriptions for insulin or oral antidiabetic medications and no records of diabetes mellitus. Matches were required to have been registered for at least 1 year before the UTS date of the same GP practice as the case at the index date of the case.

The study population was then restricted to patients from practices who participated in the linkage programme. Patients from linked practices have previously been shown to be representative of the whole CPRD population.[24] The study period was restricted to the overlapping coverage period of active follow-up in linked CPRD primary care, HES APC, cancer registration data and mortality data from the Office of National Statistics (April 1997 to December 2006) as recommended following previous research.[25] Follow-up started at the latest of the patient's index date and the start of the study period. Follow-up ended at the earliest of when a patient left the practice, the date CPRD last collected data from the practice and the end of the study period. Figure 1 shows the temporal relationship between cohort defining events, the index date and the outcome ascertainment period.

Cancer outcome ascertainment

Coded records of cancer were identified in CPRD primary care, HES APC, cancer registry and death certificate data independently. International Classification of



Time A: Maximum of: Patient registration date with the practice
Practice UTS date

Time B: Maximum of: Time A plus one year
Year in which patient turned 40 years of age
Prescription of insulin or oral antidiabetic drug
Time B is referred to as the index date and used for case/control matching

Time C: Maximum of: Time B
Start of the study period (April 1997)
Time C is the start of the outcome ascertainment period

Time D: Minimum of: End of the study period (December 2006)
Date of the patient's transfer out from the practice
Last date of data collection from the practice
Time D is the end of the outcome ascertainment period

Figure 1 Temporal relationship between cohort defining events.

Diseases 10th Revision (ICD-10) codes were used to identify cancer across HES APC, cancer registry and death certificate data (with ICD-9 being used for deaths prior to 2001), with diagnoses in primary care being made and identified using Read codes. Site-specific cancers were classified as follows: oral cavity (ICD-10 C00–14), oesophagus (C15), stomach (C16), colorectal (C18–21), pancreas (C25), head and neck (C30–32), bronchus and lung (C34), melanoma of skin (C43), non-melanoma skin cancer (NMSC) (C44), breast (C50), cervix uteri (C53), ovary (C56), prostate (C61), testis (C62), urinary organs (C64–68), brain (C71), lymphoma (C81–85), multiple myeloma (C90) or leukaemia (C91–95).

For each case recorded in CPRD primary care, it was evaluated whether HES APC contained a cancer record coded at any time, and if so, if it was of the same site. For each case recorded in HES APC, it was evaluated whether CPRD primary care contained a cancer record at any time. Coded records were searched using lists of Read code used to identify cancer outcomes in a previously published drug safety study.[23] If no coded record was found, the free text was searched for the following strings: carc, cancer, malign, chemoth, cytostat, oncolo, melanoma, lymphoma, leukaem, sarcom, myelom and metast. Records with a negative, such as 'cancer ruled out', were excluded. If a coded or anonymised free-text record of cancer was found, it was determined whether it was of the same site as identified in HES APC. For non-concordant cases recorded either in CPRD primary care or HES APC alone, cancer registry and death certificate data were reviewed for supporting evidence, such as registration of cancer in the cancer registry or mention of cancer anywhere on the death certificate.

The difference in time between cancer records of the same type in the two datasets was also evaluated by comparing the recorded dates of incident cancer cases.

Characteristics of missed cases

Variables potentially associated with non-concordance between CPRD primary care and HES APC records of cancer were evaluated using multivariable logistic regression. ORs and 95% CIs were estimated for age (calculated as year of start of follow-up minus year of birth and categorised into 40–64 (reference), 65–74 and 75+), gender (females (reference) and males) and history of type 2 diabetes (patients without diabetes (reference) and patients with type 2 diabetes). Models were fitted including all three variables (age, gender and history of type 2 diabetes).

Comparison of RRs

Finally, we used multivariable Poisson regression to estimate the RRs of cancer in patients with type 2 diabetes (as defined by primary care data) compared with patients without diabetes mellitus in each dataset. The objective of this analysis was to compare the RRs when cancer diagnoses were sourced from either primary care or hospital admissions data alone. These models also included covariates sourced from: (1) primary care data: age, gender, year of start of follow-up, smoking status, use of alcohol, body mass index and prescribing in the 6 months prior to

the start of follow-up of angiotensin II receptor blockers, antiplatelets, beta blockers, calcium channel blockers, diuretics, nitrates, non-steroidal anti-inflammatory drugs or aspirin and statins (2) linked socioeconomic status data (measured using the quintile of the Index of Multiple Deprivation[26]) and (3) primary care and/or HES APCS data: medical history of coronary heart disease, coronary revascularisation, hyperlipidaemia, hypertension, peripheral vascular disease, renal impairment and stable angina. A missing data category was used for smoking status, use of alcohol and body mass index.

Reporting

The Strengthening the Reporting of Observational Studies in Epidemiology guidelines were used to ensure the reporting of this observational study.[27]

Patient and public involvement

This study uses data provided by patients and collected by the NHS as part of their care and support. #datasaveslives

RESULTS

The study population included 53 585 patients with type 2 diabetes matched to 47 435 patients with no record of diabetes mellitus, resulting in a total study population of 101 020 patients (table 1). Just over half (53% (53 672/101 020)) were male, 45% (45 243/101 020) were aged 40–64, 30% (30 348/101 020) were between 65 and 74 years and 25% (25 439/101 020) were over 75 years.

As shown in table 2, 5797 patients had a coded record of cancer (excluding NMSC) in CPRD primary care. Of these cases, 83% (4835/5797) patients with a coded cancer record in primary care also had a record of cancer recorded in HES APC, with 78% (4542/5797) having the same site recorded in both data sources. The lowest level of concordance (43% (702/1106)) was found for NMSC, but all other concordance rates were 75% or above. Of the cases recorded in CPRD but not in HES APC, 56% (543/962) were present in either the cancer registry or death certificate data. Of the 318 cases recorded in HES APC but not in CPRD, 87% (278/318) were recorded in these other two datasets. Of the HES APC cases, 94% (5239/5557) were recorded in CPRD, 79% (4389/5557) indicating the same type of cancer and 11% (603/5557) mentioned in free-text alone.

Table 3 shows the difference in time between cancer records of the same type in CPRD primary care and HES APC data. The majority of cases were recorded within 1 month of each other. For HES APC cases, 61% (2673/4389) were recorded within 1 month in primary care and 83% (3641/4389) within 3 months. A total of 8% (382/4542) of the CPRD cases were recorded more than 1 year before the first HES APC record, whereas only 3% (128/4389) of cases were first recorded in CPRD more than 1 year after the first HES APC record.

Age was found to be positively associated with non-concordance of cancer recording (table 4). For cases recorded in HES APC, the OR for non-concordance

Table 1 Baseline characteristics of patients with type 2 diabetes and control patients without diabetes mellitus (n=101 020)

Characteristic	Type 2 diabetes (n=53 585)	Matched controls (n=47 435)
Male	28 908 (54%)	24 764 (52%)
Age, by category		
40–64 years	24 912 (46%)	20 331 (43%)
65–74 years	15 793 (30%)	14 555 (31%)
75+ years	12 880 (24%)	12 549 (26%)
Body mass index		
Underweight (<20 kg/m^2)	915 (2%)	2399 (5%)
Normal (20–25 kg/m^2)	8764 (16%)	15 012 (32%)
Overweight (25–30 kg/m^2)	19 360 (36%)	16 209 (34%)
Obese (>30 kg/m^2)	22 175 (41%)	7292 (15%)
Unknown	2371 (4%)	6523 (14%)
Smoking status		
Non smoker	23 031 (43%)	21 156 (45%)
Past smoker	16 135 (30%)	9738 (21%)
Smoker	9488 (18%)	9173 (19%)
Unknown	4931 (9%)	7368 (16%)
History of:		
Heart failure	3068 (6%)	1539 (3%)
Stable angina pectoris	7333 (14%)	4174 (9%)
Coronary heart disease	8407 (16%)	4480 (9%)
Hyperlipidaemia	3720 (7%)	1248 (3%)
Coronary revascularisation	1797 (3%)	842 (2%)
Hypertension	35 325 (66%)	19 347 (41%)
Renal impairment	1051 (2%)	486 (1%)
Peripheral vascular disease	2707 (5%)	1632 (3%)
Recent prescribing		
Organic nitrates	5840 (11%)	2585 (5%)
Beta blockers	11 786 (22%)	6137 (13%)
Calcium channel blockers	11 774 (22%)	5216 (11%)
Diuretics	18 134 (34%)	9640 (20%)
Antiplatelets	16 980 (32%)	7021 (15%)
ACE inhibitors/ angiotensin II receptor blockers	18 748 (35%)	5623 (12%)
Statins or fibrates	17 797 (33%)	4513 (10%)
Non-steroidal anti-inflammatory drugs	22 415 (42%)	12 400 (26%)

Table 2 Cancer recording across various data sources

	CPRD primary care				HES APC				
	Total no of coded cases in CPRD	No of CPRD coded cases in HES APC n (%)*	No of CPRD coded cases in HES APC with same site n (%)*	No of CPRD coded cases not in HES APC but in other data source n (%)*†	Total no of cases in HES APC	No of HES APC cases in CPRD codes or free text n (%)†	No of HES APC cases in CPRD codes or free text with same site n (%)†	No of HES APC cases in CPRD free-text alone n (%)†	No of HES APC cases not in CPRD but in other data source n (%)†‡
Any cancer (excluding NMSC)	5797	4835 (83)	4542 (78)	543 (9)	5557	5239 (94)	4389 (79)	603 (11)	278 (5)
Stomach	248	241 (97)	138 (56)	7 (3)	229	217 (95)	139 (61)	24 (11)	11 (5)
Colorectal	681	639 (94)	616 (91)	26 (4)	852	819 (96)	617 (72)	92 (11)	31 (4)
Pancreas	176	156 (89)	139 (79)	16 (9)	262	246 (94)	140 (53)	37 (14)	16 (6)
Lung	739	682 (92)	581 (79)	51 (7)	842	777 (92)	578 (69)	103 (12)	61 (7)
NMSC	1106	702 (43)	504 (31)	344 (21)	713	679 (95)	459 (64)	87 (12)	22 (3)
Breast	560	474 (85)	432 (77)	63 (11)	499	487 (98)	419 (84)	13 (3)	9 (2)
Prostate	725	542 (75)	517 (71)	122 (17)	593	574 (97)	447 (75)	37 (6)	14 (2)
Urinary organs	352	339 (96)	319 (91)	7 (2)	595	565 (95)	322 (54)	72 (12)	21 (4)
Lymphoma	201	182 (91)	166 (83)	8 (4)	203	197 (97)	164 (81)	22 (11)	5 (3)
Leukaemia	148	120 (81)	105 (71)	10 (7)	125	115 (92)	86 (69)	16 (13)	7 (6)

*Percentages calculated using number of cases identified in CPRD primary care as a denominator.
†Other data sources include cancer registration and ONS mortality data.
‡Percentages calculated using number of cases identified in HES APC as a denominator.
CPRD, Clinical Practice Research Datalink; HES APC, Hospital Episode Statistics Admitted Patient Care; NMSC, non-melanoma skin cancer; ONS, Office for National Statistics.

with CPRD primary care was more than doubled (OR 2.2; (95% CI 1.5 to 3.2)) for patients aged 75+ compared with patients aged 40–64 years. Cases aged 75+ recorded in CPRD had a 1.6-fold increased risk of non-concordance with HES APC versus patients aged 40–64 years (OR 1.6; (95% CI 1.3 to 2.1)) for patients aged 75+.

The RRs of cancer for patients with type 2 diabetes compared with matched patients without diabetes mellitus, as recorded in CPRD primary care and HES APC, are shown in table 5. The adjusted RRs were 0.90 (95% CI 0.86 to 0.96) for cancer recorded in CPRD primary care and 0.93 (95% CI 0.88 to 0.99) for cancer recorded in HES APC. Results for all cancer types were similar for outcomes recorded in CPRD and HES APC. CIs overlapped in all cases, and contained the RR estimated from the comparator source for all cancers apart from NMSC (adjusted RR 0.76 (95% CI 0.68 to 0.84) for NMSC recorded in CPRD primary care and 0.87 (95% CI 0.74 to 1.01) for NMSC recorded in HES APC).

DISCUSSION

The results of this study showed a good level of concordance in cancer recording between CPRD primary care and HES APC data, overall, in relation to the timing of the first record and in patients aged less than 75 years. The comparisons of cancer outcomes among patients with type 2 diabetes and matched patients without diabetes mellitus showed similar RRs reported in each of the two EHR databases. Together with the high level of supporting evidence for non-concordant cases from the cancer registry and death certificate data, these results suggest that misclassification of cancer in both data sources is low, except for NMSC, as expected.

However, concordance was lower when restricted to looking for recording of cancer using the same site. This was largely due to the use of non-specific cancer Read codes in both primary care and hospital admissions data, which would lead to underestimates of the incidence of site-specific cancers if either data source was used in isolation. Concordance was also lower in patients aged

Table 3 Difference in time between same-site records of cancer (excluding NMSC) in CPRD primary care and HES APC

Reference source	Comparator source	Recorded within 1 month n (%)	Recorded within 1–3 months n (%)	Recorded within 4–12 months, first in reference source n (%)	Recorded within 4–12 months, last in reference source n (%)	Recorded >1 year apart, first in reference source n (%)	Recorded >1 year apart, last in reference source n (%)
CPRD primary care	HES APC	2670 (59)	966 (21)	275(6)	162 (4)	382 (8)	87 (2)
HES APC	CPRD primary care	2673 (61)	968 (22)	174 (4)	246 (6)	128 (3)	200 (5)

CPRD, Clinical Practice Research Datalink; HES APC, Hospital Episode Statistics Admitted Patient Care; NMSC, non-melanoma skin cancer

Table 4 Variables associated with non-concordance of recording of cancer (excluding NMSC)

Source of case	Comparator source	Variable	Adjusted OR (95% CI)*
HES APC	CPRD	Aged 40–64 years	Reference
		Aged 65–74 years	1.4 (0.9 to 2.1)
		Aged 75+ years	2.2 (1.5 to 3.2)
		Females	Reference
		Males	1.0 (0.8 to 1.3)
		Matched patients without diabetes	Reference
		Patients with type 2 diabetes	0.9 (0.7 to 1.1)
CPRD	HES APC	Aged 40–64 years	Reference
		Aged 65–74 years	1.2 (0.9 to 1.5)
		Aged 75+ years	1.6 (1.3 to 2.1)
		Females	Reference
		Males	1.0 (0.8 to 1.2)
		Matched patients without diabetes	Reference
		Patients with type 2 diabetes	1.1 (0.9 to 1.3)

*Models were fitted including all three variables (age, gender and history of type 2 diabetes).
CPRD, Clinical Practice Research Datalink; HES APC, Hospital Episode Statistics Admitted Patient Care; NMSC, non-melanoma skin cancer.

75 and over. This may reflect cases where the patient died shortly after a hospital diagnosis, and information was either not sent back to the GP or was not recorded in the primary care record, or alternatively where the patient died without being hospitalised for their cancer. In addition, over 10% of cases identified in HES APC were only found in the free-text primary care records. Increased data governance regulations have subsequently led to CPRD withdrawing their provision of free-text data recorded in primary care in order to further protect patient anonymity (effective April 2016). Without these free-text data available, linked HES APC data can again reduce the risk of misclassification and underestimates of cancer incidence. Due to the positive association seen between age and non-concordance, studies focusing on older age groups may especially benefit from using linked data to capture cancer outcomes.

Few studies have been conducted comparing the recording of cancer in primary care and hospital admissions data. In the UK, a recent study considered the validity and completeness of colorectal cancer diagnoses in an alternative source of primary care data compared with HES APC in a later time period (2000–2011).[28] While this study used the alternative methodology of positive predictive values, the conclusions for colorectal cancer were similar, with a recorded positive predictive value of 98% compared with a concordance of 91% reported here. However, one of the strengths of this study was the ability to look across all cancer sites, including NMSC.

This study was limited by the challenges involved in directly comparing different EHR data sources. By their nature, primary care and hospital admissions data are sourced from different sectors of the healthcare system, with data collected for different purposes, at different frequencies and using different coding systems. It has been reported that clinical experts can disagree on the code lists from a single dictionary and therefore, non-concordance may be attributed in part to the challenges in mapping different coding dictionaries.[29] While results indicate that cancer may be recorded in primary care before hospital admissions data, this may reflect GP referrals to secondary care on the basis of suspected cancer, rather than GPs recording a confirmed diagnosis earlier than other settings. The study period was limited by the coverage period of the linked cancer registry data available at the time of the study. Furthermore, linkage between CPRD and HES APC data is dependent on the accurate recording of NHS numbers. We were not able to check the error rates of recording of NHS numbers in either data source, which would have led to overestimating non-concordance. However, previous research has identified high levels of completeness and validity of NHS numbers across primary and secondary care.[30] As this study was based on a cohort of patients with type 2 diabetes and matched patients without diabetes mellitus, the results may not be comparable to the general population. Patients with type 2 diabetes have more contacts with health services and cancer recording may be more up to date and accurate. However, we did not find major differences in cancer recording between the cases and their matched controls. It should also be noted that there are some differences between the RRs of cancer found in this study and those reported in previous meta-analyses (eg, this study shows an overall reduced risk of cancer among patients with type 2 diabetes in contrast to an increased risk reported previously).[2] As this analysis was undertaken to compare the RR when cancer diagnoses were sourced from one data source alone, rather than to best estimate the RR using all available data sources, further research using linked data to optimally define the study population, outcomes and covariates is recommended.

In conclusion, a good level of concordance in cancer recording was found between CPRD primary care and

Table 5 RRs of different types of cancer in patients with and without type 2 diabetes, by data source

Type of cancer	Source	No of cancer cases in patients with type 2 diabetes	Incidence rate/1000 person-years	No of cancer cases in matched patients without diabetes mellitus	Incidence rate/1000 person-years	Age-adjusted, sex-adjusted, calendar year-adjusted RR (95% CI)	Fully adjusted RR* (95% CI)
Any cancer (excluding NMSC)	CPRD	3073	1.61	3077	1.91	0.89 (0.84 to 0.93)	0.90 (0.86 to 0.96)
	HES APC	2891	1.52	2893	1.79	0.89 (0.85 to 0.94)	0.93 (0.88 to 0.99)
Stomach	CPRD	122	0.06	133	0.08	0.81 (0.63 to 1.04)	0.83 (0.63 to 1.09)
	HES APC	111	0.06	127	0.08	0.79 (0.61 to 1.02)	0.87 (0.65 to 1.15)
Colorectal	CPRD	377	0.19	338	0.20	1.00 (0.86 to 1.16)	1.04 (0.88 to 1.22)
	HES APC	466	0.24	417	0.25	1.01 (0.88 to 1.15)	1.05 (0.91 to 1.22)
Pancreas	CPRD	133	0.07	56	0.03	2.16 (1.58 to 2.96)	3.08 (2.19 to 4.33)
	HES APC	201	0.10	81	0.05	2.27 (1.75 to 2.94)	3.17 (2.40 to 4.20)
Lung	CPRD	333	0.17	449	0.27	0.66 (0.57 to 0.76)	0.74 (0.63 to 0.87)
	HES APC	368	0.19	509	0.31	0.64 (0.56 to 0.74)	0.74 (0.64 to 0.86)
NMSC	CPRD	811	0.42	935	0.57	0.79 (0.72 to 0.87)	0.76 (0.68 to 0.84)
	HES APC	376	0.19	378	0.23	0.92 (0.79 to 1.06)	0.87 (0.74 to 1.01)
Breast	CPRD	275	0.31	313	0.39	0.80 (0.68 to 0.94)	0.79 (0.66 to 0.96)
	HES APC	251	0.28	263	0.33	0.87 (0.73 to 1.03)	0.83 (0.68 to 1.01)
Prostate	CPRD	357	0.34	408	0.48	0.78 (0.68 to 0.90)	0.71 (0.61 to 0.83)
	HES APC	305	0.29	336	0.39	0.82 (0.70 to 0.96)	0.79 (0.66 to 0.93)
Urinary	CPRD	207	0.11	172	0.10	1.08 (0.88 to 1.32)	1.07 (0.85 to 1.34)
	HES APC	338	0.17	291	0.18	1.04 (0.89 to 1.22)	1.01 (0.85 to 1.21)
Lymphoma	CPRD	105	0.05	105	0.06	0.88 (0.67 to 1.15)	0.88 (0.65 to 1.19)
	HES APC	116	0.06	94	0.06	1.09 (0.83 to 1.43)	1.13 (0.83 to 1.54)
Leukaemia	CPRD	84	0.04	77	0.05	0.99 (0.72 to 1.34)	1.04 (0.74 to 1.46)
	HES APC	63	0.03	70	0.04	0.80 (0.57 to 1.13)	0.95 (0.65 to 1.39)

*Fully adjusted for age, sex, year of start of follow-up, smoking status, use of alcohol, body mass index, prescribing in the 6 months prior to the start of follow-up (angiotensin II receptor blockers, antiplatelets, beta blockers, calcium channel blockers, diuretics, nitrates, NSAIDs or aspirin and statins), Index of Multiple Deprivation and medical history (coronary heart disease, coronary revascularisation, hyperlipidaemia, hypertension, peripheral vascular disease, renal impairment and stable angina).
CPRD, Clinical Practice Research Datalink; HES APC, Hospital Episode Statistics Admitted Patient Care; NMSC, non-melanoma skin cancer; NSAIDs, non-steroidal anti-inflammatory drugs; RRs, relative rates.

HES APC data among patients with type 2 diabetes and matched controls. However, when analysis is focused on site-specific cancers, linked data have the potential to reduce misclassification and increase case ascertainment over using either data source in isolation.

Acknowledgements The authors thank Dr Puja Myles and Dr Lucy Carty (CPRD) for their comments on an earlier version of this manuscript.

Contributors RW, T-PvS, AG and TH designed the study. RW drafted the manuscript. RW, T-PvS, AG, TH, HGML and FdV contributed to the interpretation of the results and revision of the manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests All authors confirm that they are not involved in any organisation or entity with a financial interest in or financial conflict with the subject matter or materials discussed in this manuscript. CPRD is owned by the Secretary of State of the UK Department of Health and operates within the MHRA. CPRD has received funding from the MHRA, Wellcome Trust, Medical Research Council, NIHR Health Technology Assessment programme, Innovative Medicine Initiative, UK Department of Health, Technology Strategy Board, Seventh Framework Programme EU, various universities, contract research organisations and pharmaceutical companies. The Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute for Pharmaceutical Sciences has received unrestricted funding for pharmacoepidemiological research from GlaxoSmithKline, Novo Nordisk, the private-public funded Top Institute Pharma (www.tipharma.nl, includes cofunding from universities, government and industry), the Dutch Medicines Evaluation Board and the Dutch Ministry of Health.

Patient consent Not required.

Ethics approval Ethical approval for all purely observational research using anonymised CPRD data has been obtained from the East Midlands—Derby Research Ethics Service Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data for this study were derived from CPRD primary care and linked data obtained under license from the UK Medicines and Healthcare products Regulatory Agency. While these data cannot be shared, in order to respect the wishes of patients who have opted out or dissented since the study was conducted, data for similar cohorts are available from CPRD subject to protocol approval and license agreements.

REFERENCES

1. International Diabetes Federation. *IDF Diabetes Atlas*. 7th edn. Brussels, Belgium: International Diabetes Federation, 2015.
2. Tsilidis KK, Kasimis JC, Lopez DS, *et al*. Type 2 diabetes and cancer: umbrella review of meta-analyses of observational studies. *BMJ* 2015;350:g7607.
3. Ehrenstein V, Petersen I, Smeeth L, *et al*. Helping everyone do better: a call for validation studies of routinely recorded health data. *Clin Epidemiol* 2016;8:49–51.
4. Bernal-Delgado EE, Martos C, Martínez N, *et al*. Is hospital discharge administrative data an appropriate source of information for cancer registries purposes? Some insights from four Spanish registries. *BMC Health Serv Res* 2010;10:9.
5. Creighton N, Walton R, Roder D, *et al*. Validation of administrative hospital data for identifying incident pancreatic and periampullary cancer cases: a population-based study using linked cancer registry and administrative hospital data in New South Wales, Australia. *BMJ Open* 2016;6:e011161.
6. Funch D, Ross D, Gardstein BM, *et al*. Performance of claims-based algorithms for identifying incident thyroid cancer in commercial health plan enrollees receiving antidiabetic drug therapies. *BMC Health Serv Res* 2017;17:330.
7. Goldsbury D, Weber M, Yap S, *et al*. Identifying incident colorectal and lung cancer cases in health service utilisation databases in Australia: a validation study. *BMC Med Inform Decis Mak* 2017;17:23.
8. Kemp A, Preen DB, Saunders C, *et al*. Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia. *BMC Med Res Methodol* 2013;13:17.
9. Leinonen MK, Miettinen J, Heikkinen S, *et al*. Quality measures of the population-based Finnish Cancer Registry indicate sound data quality for solid malignant tumours. *Eur J Cancer* 2017;77:31–9.
10. Palmaro A, Gauthier M, Conte C, *et al*. Identifying multiple myeloma patients using data from the French health insurance databases. *Medicine* 2017;96:e6189.
11. Seo HJ, Oh IH, Yoon SJ. A comparison of the cancer incidence rates between the national cancer registry and insurance claims data in Korea. *Asian Pac J Cancer Prev* 2012;13:6163–8.
12. 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.
13. Boggon R, van Staa TP, Chapman M, *et al*. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiol Drug Saf* 2013;22:168–75.
14. Dregan A, Moller H, Murray-Thomas T, *et al*. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer Epidemiol* 2012;36:425–9.
15. Sollie A, Roskam J, Sijmons RH, *et al*. Do GPs know their patients with cancer? Assessing the quality of cancer registration in Dutch primary care: a cross-sectional validation study. *BMJ Open* 2016;6:e012669.
16. CPRD. Cancer Registration Data and GOLD Documentation (Set 13). *Unpublished documentation* 2016.
17. Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
18. Williams T, van Staa T, Puri S, *et al*. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv Drug Saf* 2012;3:89–99.
19. 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.
20. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010;60:128–36.
21. Digital NHS. Hospital Episode Statistics. http://content.digital.nhs.uk/hes (accessed 3 Jun 2017).
22. Digital NHS. Data quality checks performed on SUS and HES data. http://content.digital.nhs.uk/media/13655/Data-quality-checks-performed-on-SUS-and-HES-data/pdf/HESDQ_In_002_Data_quality_checks_performed_on_SUS_and_HES_data.pdf (accessed 28 Jan 2018).
23. van Staa TP, Patel D, Gallagher AM, *et al*. Oral antidiabetics and insulin and the patterns of risk of cancer: a study with the General Practice Research Database and secondary care data. *Diabetologia* 2012;55:654–65.
24. Gallagher AM, Puri D, van Staa TP. Linkage of the General Practice Research Database (GPRD) with other data sources. *Pharmacoepidemiology & Drug Safety* 2011;20(S1):S364.
25. Gallagher AM, Williams T, Leufkens HG, *et al*. The Impact of the Choice of Data Source in Record Linkage Studies Estimating Mortality in Venous Thromboembolism. *PLoS One* 2016;11:e0148349.
26. Communities and Local Government. English indices of deprivation 2010. http://www.gov.uk/government/statistics/english-indices-of-deprivation-2010 (accessed 24 Jun 2017).
27. University of Bern. STROBE Statement. http://www.strobe-statement.org (accessed 26 Nov 2017).
28. Cea Soriano L, Soriano-Gabarró M, García Rodríguez LA. Validity and completeness of colorectal cancer diagnoses in a primary care database in the United Kingdom. *Pharmacoepidemiol Drug Saf* 2016;25:385–91.
29. Gulliford MC, Charlton J, Ashworth M, *et al*. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. *PLoS One* 2009;4:e7168.
30. Hippisley-Cox J. *Validity and completeness of the NHS Number in primary and secondary care: electronic data in England 1991-2013*. Nottingham: University of Nottingham, 2013.