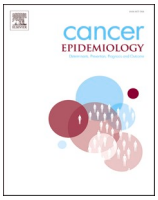
Cancer Epidemiology 94 (2025) 102715

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep



Evaluating the quality of prostate cancer diagnosis recording in CPRD GOLD and CPRD Aurum primary care databases for observational research: A study using linked English electronic health records

Gayasha Somathilake [a,*,1], Elizabeth Ford [b], Jo Armes [a], Sotiris Moschoyiannis [c], Michelle Collins [d], Patrick Francsics [d], Agnieszka Lemanska [a,e]

[a] School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, UK
[b] Department of Primary Care and Public Health, Brighton and Sussex Medical School, UK
[c] Computer Science Research Centre, Faculty of Engineering and Physical Sciences, University of Surrey, UK
[d] Royal Surrey County Hospital, Guildford, UK
[e] Data Science, National Physical Laboratory, Teddington, UK

ARTICLE INFO

Keywords:
CPRD GOLD
CPRD Aurum
Data linkage
Data quality
Validation

ABSTRACT

Background: Primary care data in the UK are widely used for cancer research, but the reliability of recording key events like diagnoses remains uncertain. Although data linkage can improve reliability, its costs, time requirements, and sample size constraints may discourage its use. We evaluated accuracy, completeness, and date concordance of prostate cancer (PCa) diagnosis recording in Clinical Practice Research Datalink (CPRD) GOLD and Aurum compared to linked Cancer Registry (CR) and Hospital Episode Statistics (HES) Admitted Patient Care (APC) in England.

Methods: Incident PCa diagnoses (2000–2016) for males aged ≥46 at diagnosis who remained registered with their General Practitioner (GP) by age 65 and were recorded in at least one data source were analysed. Accuracy was the proportion of diagnoses recorded in GOLD or Aurum with a corresponding record in CR or HES. Completeness was the proportion of CR or HES diagnoses with a corresponding record in GOLD or Aurum.

Results: The final cohorts for comparisons included 29,500 records for GOLD and 26,475 for Aurum. Compared to CR, GOLD was 86 % accurate and 65 % complete, while Aurum was 87 % accurate and 77 % complete. Compared to HES, GOLD was 76 % accurate and 60 % complete, and Aurum was 79 % accurate and 70 % complete. Concordance in diagnosis dates improved over time in both GOLD and Aurum, with 93 % of diagnoses recorded within a year compared to CR, and 66 % (GOLD) and 71 % (Aurum) compared to HES. Delays of 2–3 weeks in primary care diagnosis recording were observed compared to CR, whereas most diagnoses appeared at least 3 months earlier in primary care than in HES.

Conclusions: Aurum demonstrated better accuracy and completeness for PCa diagnosis recording than GOLD. However, linkage to HES or CR is recommended for improved case capture. Researchers should address the limitations of each data source to ensure research validity.

1. Introduction

In the United Kingdom (UK), primary care electronic health record (EHR) databases draw data from National Health Service (NHS) general practices where over 98 % of the population are registered to access primary healthcare [1]. These databases are extensively used for cancer research since most patients with symptoms initially present to a general practitioner (GP) [2–6]. However, identifying cancer diagnoses from primary care EHRs presents challenges due to inaccurate and incomplete data varying across tumour types [7–9]. Additionally, a diagnostic code in primary care may not definitively indicate a cancer diagnosis, while lacking a code may not consistently imply the absence. Recent improvements in data linkage provide solutions to mitigate these uncertainties and the potential misclassification bias when studies solely

* Corresponding author.
E-mail address: g.somathilake@surrey.ac.uk (G. Somathilake).
[1] ORCID: 0000-0002-1689-3881.

https://doi.org/10.1016/j.canep.2024.102715
Received 30 July 2024; Received in revised form 8 November 2024; Accepted 17 November 2024
Available online 30 November 2024

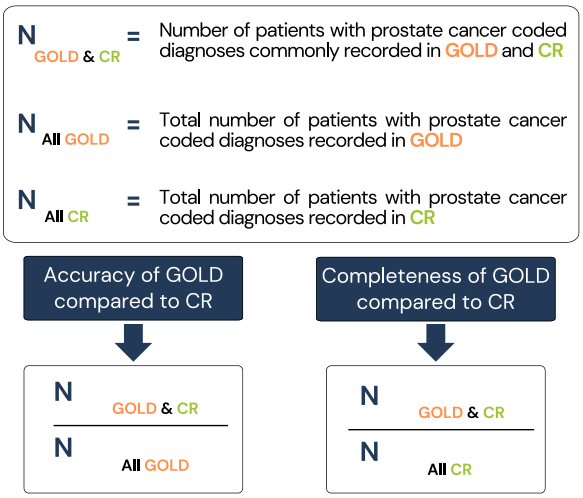


Fig. 1. Definitions of accuracy and completeness for GOLD compared to CR for reference.

rely on primary care data [10]. Nonetheless, researchers might still avoid data linkage due to high costs, tight timelines, and limitations in sample size and coverage [11]. This underscores the importance of assessing the correctness, completeness, and timeliness of primary care data for research purposes [12].

Prostate cancer (PCa) is the most common malignancy among men in the UK typically diagnosed with GP referrals following abnormal exams or elevated prostate-specific antigen (PSA) levels. [13,14]. For diagnosis information to appear in primary care records, hospital discharge or diagnosis letters must be accurately coded into EHRs. This can cause delays, as primary care diagnosis records are often only updated when GPs take action, such as prescribing medication [15]. Moreover, errors in data transfer, or misclassification, such as confusing benign prostatic hypertrophy/hyperplasia (BPH) with PCa make identifying PCa diagnoses from primary care EHRs challenging. Hence, this study aimed to evaluate the quality of PCa diagnosis recording in primary care using linked Clinical Practice Research Datalink (CPRD) data in England.

CPRD GOLD and Aurum are UK primary care datasets providing patient-level data linkage to the National Cancer Registration and Analysis Service (NCRAS), Hospital Episode Statistics (HES), and Office of National Statistics (ONS) in England [1,16–18]. This study evaluated the quality of PCa diagnoses in GOLD and Aurum based on their accuracy, completeness and recording dates through comparison to external datasets: NCRAS Cancer Registry (CR) and HES Admitted Patient Care (APC). We also explored patient records with diagnoses coded in both linked sources but without corresponding records in GOLD and Aurum. The findings offer insights into primary care data quality to support informed decisions in selecting data sources for PCa research.

2. Methods

2.1. Data sources

This study utilised CPRD GOLD and CPRD Aurum primary care data, linked to the Cancer Registry (CR), Hospital Episode Statistics Admitted Patient Care (HES APC), and the 2015 Index of Multiple Deprivation (IMD) quintiles [19] (November 2019 release version, set 17). We will refer to CPRD GOLD as 'GOLD', CPRD Aurum as 'Aurum', Cancer Registry as 'CR,' and HES APC data as 'HES.'

GOLD and Aurum are two large, longitudinal, UK population-based EHR databases widely used for medical research. Both databases collect deidentified patient data from GPs, including demographic

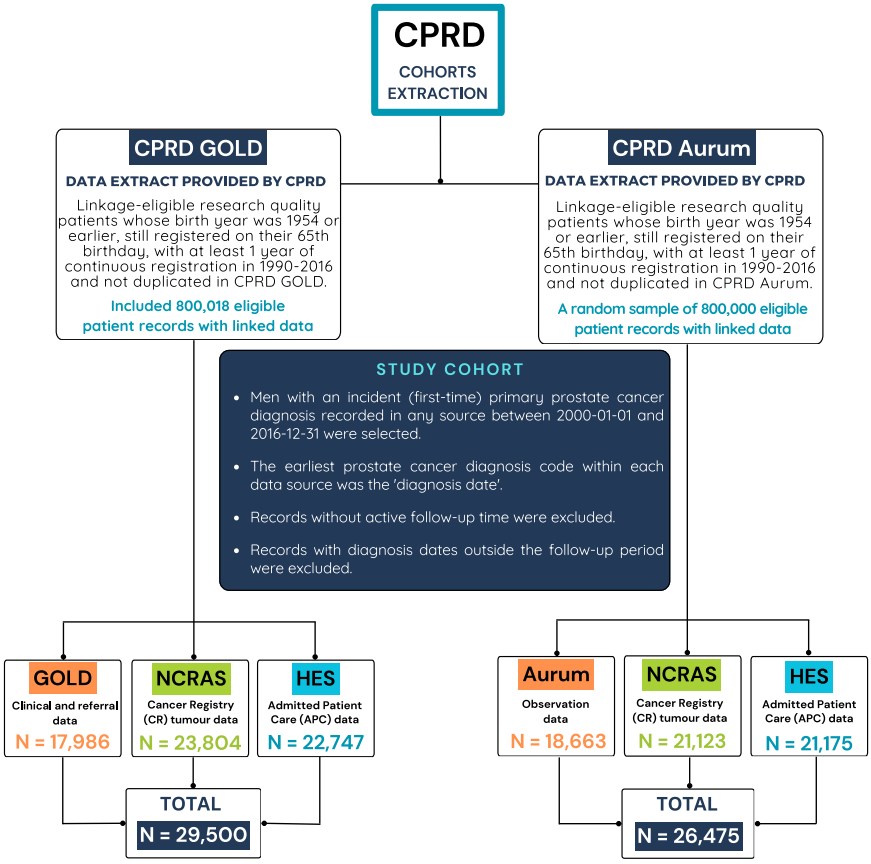


Fig. 2. Cohort definitions for CPRD GOLD and CPRD Aurum and the final study cohorts.

Table 1
Patient characteristics between study data sources in each cohort.

Characteristics	CPRD GOLD cohort (N = 29,500)			CPRD Aurum cohort (N = 26,475)		
	GOLD (N = 17,986)	CR (N = 23,804)	HES (N = 22,747)	Aurum (N = 18,663)	CR (N = 21,123)	HES (N = 21,175)
Age at diagnosis						
Mean ± st. dev.	73.6 ± 8.1	73.8 ± 8.1	76.1 ± 8.5	73.9 ± 8.4	73.1 ± 8.4	75.3 ± 8.9
Median (IQR)	73 [67–79]	74 [68–79]	76 [70–82]	73 [67–79]	73 [67–79]	75 [69–82]
< 60 years	545 (3 %)	680 (3 %)	405 (2 %)	828 (4 %)	928 (4 %)	633 (3 %)
60–69 years	5534 (31 %)	7074 (30 %)	5236 (23 %)	5980 (32 %)	6625 (31 %)	5366 (25 %)
70–79 years	7519 (42 %)	10,266 (43 %)	9177 (40 %)	7534 (40 %)	8654 (41 %)	8100 (38 %)
≥ 80 years	4388 (24 %)	5784 (24 %)	7929 (35 %)	4321 (23 %)	4916 (23 %)	7076 (33 %)
Year of diagnosis						
2000–2003	4061 (23 %)	4681 (20 %)	3818 (17 %)	3759 (20 %)	3984 (19 %)	3598 (17 %)
2004–2007	5061 (28 %)	5364 (23 %)	4777 (21 %)	4458 (24 %)	4766 (23 %)	4503 (21 %)
2008–2011	4874 (27 %)	6078 (26 %)	5973 (26 %)	4693 (25 %)	5323 (25 %)	5603 (26 %)
2012–2016	3990 (22 %)	7681 (32 %)	8179 (36 %)	5753 (31 %)	7050 (33 %)	7471 (35 %)
Follow-up duration (years)						
Mean ± st. dev.	10.4 ± 4.9	9.9 ± 4.9	9.4 ± 4.9	12.5 ± 5.3	12.8 ± 5.3	12.4 ± 5.4
Median (IQR)	11.0 (6.4–14.7)	10.4 (5.8–14.0)	9.8 (5.1–13.5)	15.2 (8.2–17.0)	16.1 (8.7–17.0)	15.2 (8.0–17.0)
IMD						
1 - least deprived	2719 (15 %)	3650 (15 %)	3430 (15 %)	4007 (21 %)	4441 (21 %)	4347 (21 %)
2	2934 (16 %)	3920 (16 %)	3817 (17 %)	3921 (21 %)	4358 (21 %)	4251 (20 %)
3	3873 (22 %)	5105 (21 %)	4786 (21 %)	3698 (20 %)	4172 (20 %)	4230 (20 %)
4	4132 (23 %)	5711 (24 %)	5386 (24 %)	3717 (20 %)	4228 (20 %)	4328 (20 %)
5 - most deprived	4009 (22 %)	5418 (23 %)	5328 (23 %)	3320 (18 %)	3924 (19 %)	4019 (19 %)
Missing	319 (2 %)	-	-	-	-	-
Ethnicity						
Asian	127 (1 %)	182 (1 %)	182 (1 %)	250 (1 %)	277 (1 %)	279 (1 %)
Black	279 (2 %)	370 (2 %)	350 (2 %)	561 (3 %)	626 (3 %)	612 (3 %)
Mixed	41 (0 %)	49 (0 %)	38 (0 %)	69 (0 %)	74 (0 %)	67 (0 %)
Other	90 (1 %)	131 (1 %)	115 (1 %)	105 (1 %)	123 (1 %)	128 (1 %)
White	16,374 (91 %)	22,089 (93 %)	21,426 (94 %)	17,019 (91 %)	19,274 (91 %)	19,523 (92 %)
Unknown	1075 (6 %)	983 (4 %)	636 (3 %)	659 (4 %)	749 (4 %)	566 (3 %)
Practice Region						
1-North East	447 (2 %)	654 (3 %)	644 (3 %)	755 (4 %)	831 (4 %)	874 (4 %)
2-North West	1529 (9 %)	1980 (8 %)	2058 (9 %)	2988 (16 %)	3308 (16 %)	3578 (17 %)
3-Yorkshire & The Humber	906 (5 %)	1411 (6 %)	1337 (6 %)	745 (4 %)	819 (4 %)	785 (4 %)
4-East Midlands	646 (4 %)	1101 (5 %)	1069 (5 %)	360 (2 %)	416 (2 %)	395 (2 %)
5-West Midlands	1481 (8 %)	1833 (8 %)	1838 (8 %)	3827 (21 %)	4298 (20 %)	4124 (19 %)
6-East of England	2326 (13 %)	3391 (14 %)	3320 (15 %)	1161 (6 %)	1255 (6 %)	1214 (6 %)
7-South West	3491 (19 %)	4876 (20 %)	4417 (19 %)	2916 (16 %)	3409 (16 %)	3435 (16 %)
8-South Central	2854 (16 %)	3346 (14 %)	2874 (13 %)	2257 (12 %)	2541 (12 %)	2313 (11 %)
9-London	1454 (8 %)	1884 (8 %)	1873 (8 %)	1914 (10 %)	2267 (11 %)	2404 (11 %)
10-South East Coast	2852 (16 %)	3328 (14 %)	3317 (15 %)	1692 (9 %)	1927 (9 %)	2004 (9 %)
Missing	-	-	-	48 (0 %)	52 (0 %)	49 (0 %)

details, medical diagnoses, prescriptions, referrals, and hospital information [1,18]. They offer a secure data linkage service for external healthcare and area-level databases for a subset of English general practices that have consented. This linkage is established through a unique patient identifier created by CPRD, allowing for the integration and analysis of data from multiple sources. GOLD, sourced from Vision patient management software, has been available for decades and contains data from over 15 million patients [1]. It is well-validated, with over 2400 peer-reviewed publications [9]. However, its patient numbers have declined as practices transition to other systems [1]. In contrast, Aurum, which covers over 30 million patients from more than 1000 practices, and is drawn from EMIS patient management software, has rapidly grown in usage and offers a larger, more contemporary patient population [18]. Aurum uses a combination of coding systems, including SNOMED CT [20,21] and Read Version 2, whereas GOLD primarily uses Read codes [22,23]. For this study, data from both databases covered the period from 1 January 1990–31 December 2016, marking the available endpoint at the time of download.

The CR from NCRAS is a dynamic population-based database that captures information across the entire cancer care pathway including cancer diagnoses, treatments, and outcomes and serves as the national standard for reporting cancer in England [16]. Each registerable tumour diagnosed and treated in England is documented in the CR and reported to NCRAS. Cancer diagnoses are systematically recorded using ICD-10 codes, with information sourced from various entities such as hospitals, pathology and treatment reports, hospices, as well as cancer screening and treatment centers. Over the years, CR has undergone several transformations to improve data quality, particularly after the introduction of the Cancer Outcomes and Services Dataset (COSD) in 2013, which set new standards for data collection and reporting [16]. This evolution reflects ongoing efforts to adapt to changing clinical practices and to incorporate new data sources. CR data was available from 1 January 1990–31 December 2016, the endpoint at the time of download. HES records all admissions, outpatient appointments, and Accident and Emergency attendances at NHS hospitals. HES APC includes admission and discharge dates, diagnoses (using ICD-10 codes) and procedures for inpatient hospitalisations [17]. HES data used for this study spanned from 1 April 1997–31 December 2016 the endpoint at the time of download.

2.2. Patient selection

2.2.1. Source populations

The source populations provided by the CPRD included linkage-eligible patients born in 1954 or earlier, with at least one year of continuous registration between 1 January 1990 and 31 December 2016, and who were still registered on their 65th birthday (Fig. 2). CPRD only included patients from English practices contributing data of acceptable quality and those from practices that switched from InPS Vision to EMIS were excluded to avoid duplication between GOLD and Aurum. Consequently, the GOLD population included all 800,018 patients who met the above inclusion criteria, while the Aurum population

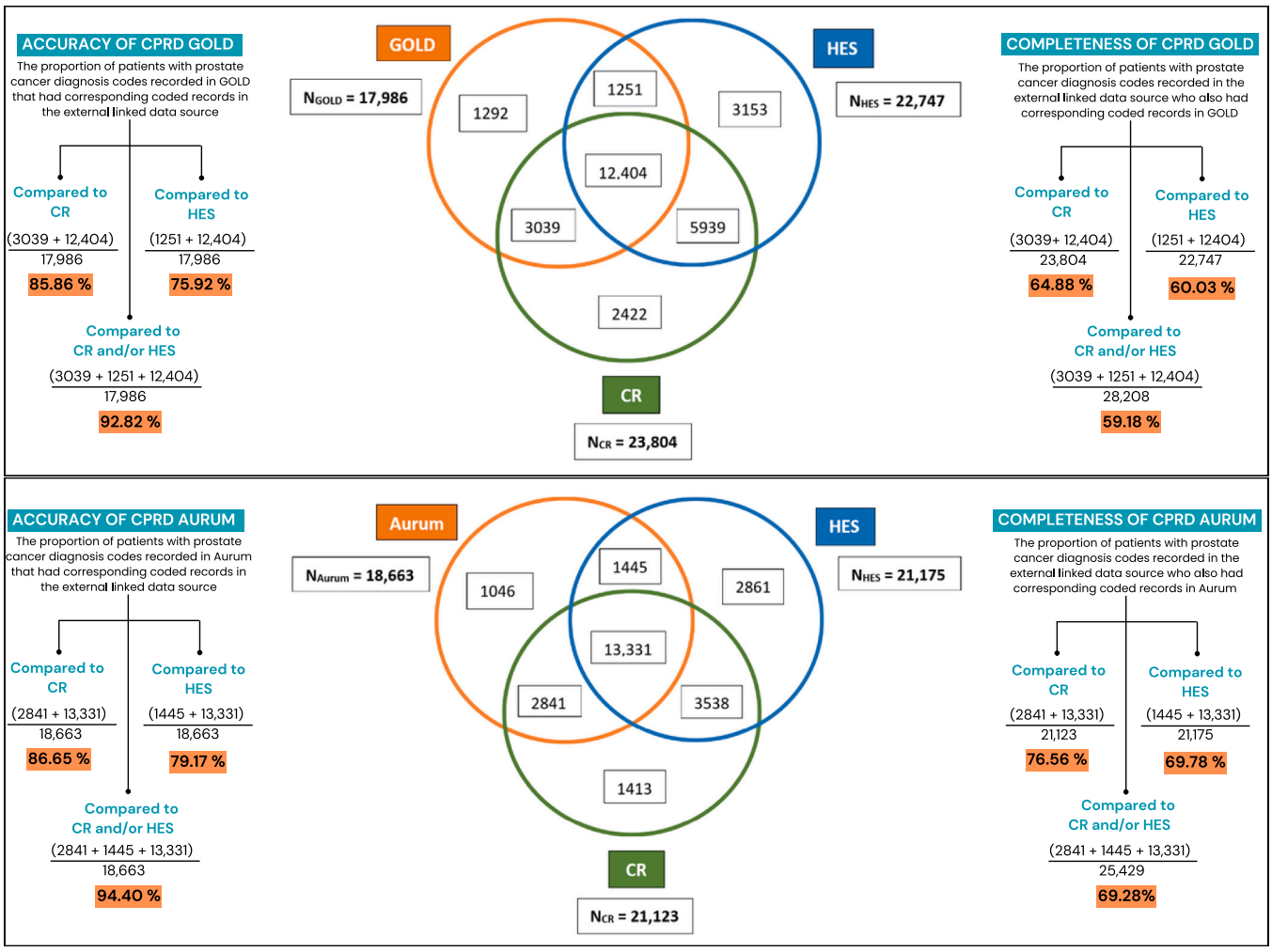


Fig. 3. The agreement between study sources and overall accuracy/completeness measures in CPRD GOLD and CPRD Aurum.

was a random sample of 800,000 eligible patients, matched by size to GOLD.

2.2.2. Study cohorts

To define our study cohorts for GOLD and Aurum with linkage to CR and HES, we extracted records of male patients who had a first-time (incident) primary PCa diagnosis recorded in at least one of the data sources between 1 January 2000 and 31 December 2016 (Supplementary_information-S1) (Fig. 2). Follow-up periods were defined for each patient record to identify the period during which all data sources were present. The start of follow-up was the latest of either the primary care registration start date/current registration date, or the beginning of the study period (1 January 2000). The end of follow-up was set to the earliest of the transfer-out date, the last data collection date, the date of death, or the end of the study period (31 December 2016). The date of the earliest PCa diagnosis code in each data source was defined as the 'diagnosis date' for each patient. Records with no active follow-up time (i.e., if end of follow-up<start of follow-up) or diagnosis dates outside the follow-up period were excluded to ensure diagnoses occurred within active follow-up.

2.3. Patient characteristics

Patient characteristics were categorised as follows: age at diagnosis (<60, 60–69, 70–79, ≥80) with a minimum diagnosis age of 46 years; year of diagnosis (2000–2003, 2004–2007, 2008–2011, 2012–2016); practice-level IMD quintile (1 = least deprived to 5 = most deprived); ethnicity (aggregated based on the higher-level ethnicity classification with six groups: Asian, Black, Mixed, Other, Unknown, White) and the geographical region of the registered practice in England.

2.4. Data analysis

The analysis was conducted separately for GOLD and Aurum, comparing each source to CR and HES. The study cohorts for GOLD and Aurum were defined, and patient characteristics were detailed alongside an assessment of the overall agreement between data sources. The quality of PCa diagnoses in GOLD and Aurum was explored through three key metrics: accuracy, completeness, and recording date concordance.

Accuracy of GOLD/Aurum was defined as the proportion of patients with PCa diagnoses in GOLD/Aurum that had corresponding records in the external source. Completeness of GOLD/Aurum was the proportion of all patients with diagnoses in the external source that had corresponding records in GOLD/Aurum. Methodologies by Weiskopf and Weng [24] were employed for these estimations, with example calculations for GOLD compared to CR illustrated in Fig. 1. These metrics were further analysed based on the patient's age at diagnosis, diagnosis year, IMD, ethnicity, and region.

Discrepancies in diagnosis dates were assessed by comparing GOLD and Aurum with both CR and HES separately. The differences in dates were calculated in weeks, with the mean, standard deviation, median, and interquartile range (IQR) used to describe each comparison. For clarity in plotting these differences, records with more than one-year

Table 2
Accuracy estimates of prostate cancer diagnoses recorded in CPRD GOLD and Aurum compared to CR and HES.

Accuracy estimates by patient characteristics compared to CR

Characteristics	CPRD GOLD			CPRD Aurum		
	Cases in GOLD (N)	Cases in GOLD and CR (N)	Accuracy Estimate (%)	Cases in Aurum (N)	Cases in Aurum and CR (N)	Accuracy Estimate (%)
Overall accuracy estimates	17,986	15,443	85.9 %	18,663	16,172	86.7 %
Age at diagnosis						
< 60 years	545	460	84.4 %	828	717	86.6 %
60–69 years	5534	4900	88.5 %	5980	5389	90.1 %
70–79 years	7519	6601	87.8 %	7534	6675	88.6 %
≥ 80 years	4388	3482	79.4 %	4321	3391	78.5 %
Year of diagnosis						
2000–2003	4061	3123	76.9 %	3759	2930	77.9 %
2004–2007	5061	4316	85.3 %	4458	3749	84.1 %
2008–2011	4874	4345	89.1 %	4693	4199	89.5 %
2012–2016	3990	3659	91.7 %	5753	5294	92.0 %
IMD						
1 - least deprived	2719	2349	86.4 %	4007	3420	85.4 %
2	2934	2518	85.8 %	3921	3412	87.0 %
3	3873	3387	87.5 %	3698	3174	85.8 %
4	4132	3597	87.1 %	3717	3216	86.5 %
5 - most deprived	4009	3592	89.6 %	3320	2950	88.9 %
Ethnicity						
Asian	127	110	86.6 %	250	216	86.4 %
Black	279	245	87.8 %	561	488	87.0 %
Mixed	41	35	85.4 %	69	64	92.8 %
Other	90	78	86.7 %	105	89	84.8 %
White	16,374	14,398	87.9 %	17,019	14,817	87.1 %
Practice Region						
1-North East	447	416	93.1 %	755	691	91.5 %
2-North West	1529	1382	90.4 %	2988	2595	86.8 %
3-Yorkshire & The Humber	906	794	87.6 %	745	662	88.9 %
4-East Midlands	646	578	89.5 %	360	304	84.4 %
5-West Midlands	1481	1259	85.0 %	3827	3374	88.2 %
6-East of England	2326	2050	88.1 %	1161	1002	86.3 %
7-South West	3491	3122	89.4 %	2916	2581	88.5 %
8-South Central	2854	2294	80.4 %	2257	1955	86.6 %
9-London	1454	1224	84.2 %	1914	1579	82.5 %
10-South East Coast	2852	2324	81.5 %	1692	1388	82.0 %

Accuracy estimates by patient characteristics compared to HES

Characteristics	CPRD GOLD			CPRD Aurum		
	Cases in GOLD (N)	Cases in GOLD and HES (N)	Accuracy Estimate (%)	Cases in Aurum (N)	Cases in Aurum and HES (N)	Accuracy Estimate (%)
Overall accuracy estimates	17,986	13,655	75.9 %	18,663	14,776	79.2 %
Age at diagnosis						
< 60 years	545	408	74.9 %	828	667	80.6 %
60–69 years	5534	4280	77.3 %	5980	4875	81.5 %
70–79 years	7519	5713	76.0 %	7534	5935	78.8 %
≥ 80 years	4388	3254	74.2 %	4321	3299	76.3 %
Year of diagnosis						
2000–2003	4061	3069	75.6 %	3759	3000	79.8 %
2004–2007	5061	3974	78.5 %	4458	3684	82.6 %
2008–2011	4874	3786	77.7 %	4693	3882	82.7 %
2012–2016	3990	2826	70.8 %	5753	4210	73.2 %
IMD						
1 - least deprived	2719	2058	75.7 %	4007	3029	75.6 %
2	2934	2242	76.4 %	3921	3041	77.6 %
3	3873	2951	76.2 %	3698	2955	79.9 %
4	4132	3165	76.6 %	3717	3004	80.8 %
5 - most deprived	4009	3239	80.8 %	3320	2747	82.7 %
Ethnicity						
Asian	127	100	78.7 %	250	197	78.8 %
Black	279	223	79.9 %	561	449	80.0 %
Mixed	41	27	65.9 %	69	51	73.9 %
Other	90	63	70.0 %	105	86	81.9 %
White	16,374	12,923	78.9 %	17,019	13,672	80.3 %
Practice Region						
1-North East	447	375	83.9 %	755	643	85.2 %
2-North West	1529	1345	88.0 %	2988	2588	86.6 %
3-Yorkshire & The Humber	906	715	78.9 %	745	586	78.7 %
4-East Midlands	646	533	82.5 %	360	279	77.5 %

(*continued on next page*)

Table 2 (*continued*)

5-West Midlands	1481	1143	77.2 %	3827	2960	77.3 %
6-East of England	2326	1881	80.9 %	1161	893	76.9 %
7-South West	3491	2680	76.8 %	2916	2466	84.6 %
8-South Central	2854	1793	62.8 %	2257	1576	69.8 %
9-London	1454	1098	75.5 %	1914	1484	77.5 %
10-South East Coast	2852	2092	73.4 %	1692	1271	75.1 %

gap between diagnosis dates were excluded. Finally, patients identified with PCa diagnoses in both CR and HES but lacking corresponding records in GOLD/Aurum were examined within ±365 days of their CR diagnosis date to identify potential reasons. The presence of supporting clinical codes related to PCa management, such as prostatectomy, radiotherapy, chemotherapy, and referrals/specialist visits, was explored restricting the analysis to men without a history of other malignancies. Additional possible explanations for variations in recording were also described.

This study adhered to the REporting of studies Conducted using the Observational Routinely collected health Data (RECORD) statement [25] for reporting observational research involving routinely collected data (Supplementary_information-S2). All statistical analyses were performed using R software (version 4.3.2).

3. Results

3.1. Study cohort definitions

The final study cohorts for comparison included 29,500 patients for GOLD and 26,475 patients for Aurum with PCa diagnosis codes recorded in any data source during the study period (Fig. 2).

Age at diagnosis was similar across GOLD, Aurum and CR, but HES included older patients in both cohorts: the mean age was 76.1 years in GOLD and 75.3 years in Aurum (Table 1). HES had a higher proportion of patients aged ≥80 years across both cohorts (35 % in GOLD and 33 % in Aurum) compared to 23–24 % in other sources. Diagnoses increased over time in all sources, except GOLD with fewer diagnoses in 2012–2016 (22 %). Aurum had more patients from less deprived areas (21 % in Aurum vs. 15 % in GOLD), while GOLD had more from the most deprived (22 % in GOLD vs. 18 % in Aurum). Ethnicity was similar across sources, with >90 % of patients being White. Aurum had more patients from practices in the West Midlands, North West, and South West (16–21 %), while GOLD had more from the South West, South Central, and South East Coast (16–19 %).

3.2. Agreement between study sources

For the comparison of GOLD with CR and HES, a total of 29,500 diagnoses were identified from either source, with 61 % recorded in GOLD, 81 % in CR, and 77 % in HES (Fig. 3). The accuracy estimates of GOLD were 85.9 % when compared to CR and 75.9 % compared to HES, while the completeness estimates were 64.9 % for CR and 60.0 % for HES. In the Aurum comparison, 26,475 diagnoses in total were identified; 71 % were recorded in Aurum, with 80 % each in CR and HES. Aurum demonstrated better quality estimates, with an accuracy of 86.7 % for CR and 79.2 % for HES, along with completeness estimates of 76.6 % and 69.8 %, respectively.

3.3. Accuracy estimates by patient characteristics

Accuracy estimates for both GOLD and Aurum compared to CR were generally similar across most patient characteristics (Table 2). The lowest accuracy was observed in the ≥80 age group for both cohorts (~79 %), while accuracy improved over time, reaching 92 % for diagnoses made between 2012 and 2016. IMD-5 quintile exhibited the highest accuracy (89–90 %). The 'mixed' ethnic group had the highest accuracy in Aurum (93 %) but the least in GOLD (85 %). Practices in the North East demonstrated the highest accuracy (92–93 %), whereas South Central had the lowest accuracy in GOLD (80 %) and South East Coast in Aurum (82 %).

Compared to HES, Aurum generally showed higher accuracy than GOLD (Table 2). Accuracy was again lowest for the ≥80-year age group (74 % in GOLD, 76 % in Aurum) and for diagnoses recorded in 2012–2016 (71 % in GOLD, 73 % in Aurum). IMD quintile 5 also had the highest accuracy (81–83 %), while the 'mixed' ethnic group had the lowest (66 % in GOLD, 74 % in Aurum). The North West region had the highest accuracy (87–88 %), while South Central had the lowest (63 % in GOLD, 70 % in Aurum).

3.4. Completeness estimates by patient characteristics

Completeness estimates compared to CR were notably higher for Aurum than for GOLD across all patient characteristics (Table 3). Completeness declined with age, with the ≥80 age group showing the lowest data completeness (56 % for GOLD and 64 % for Aurum). Completeness in GOLD was the least during 2012–2016 (46 %), while Aurum had its lowest in 2000–2003 (72 %). The 'mixed' ethnic group showed higher completeness in both cohorts (71 % in GOLD and 87 % in Aurum). The highest completeness for GOLD was seen in the North West and South East Coast (70 %), with the East Midlands having the lowest (53 %). In Aurum, the North East had the highest completeness (83 %), while London had the lowest (70 %).

Compared to HES, Aurum again demonstrated higher completeness than GOLD across all patient characteristics (Table 3). The ≥80 age group had the lowest completeness (52 % in GOLD, 57 % in Aurum). Diagnoses recorded between 2012 and 2016 were least complete in GOLD (51 %), while Aurum had its lowest in 2000–2003 (58 %). The 'mixed' ethnic group also showed higher completeness in both cohorts (71 % in GOLD, 76 % in Aurum). The highest completeness for GOLD was in the North West (65 %) and the lowest in the East Midlands (50 %). In Aurum, Yorkshire & The Humber had the highest completeness (75 %), while London had the lowest (62 %).

3.5. Discrepancies in diagnosis dates

In comparison to CR, a majority of records in both GOLD (77 %) and Aurum (74 %) showed later diagnosis dates, with median differences of 15 days in both databases (Table 4). Conversely, when compared to HES, a significant proportion of records in GOLD (62 %) and Aurum (59 %) had earlier diagnosis dates, with median differences of 67 days and 42 days, respectively. Only 6 % of records in both databases matched the diagnosis dates in CR, while 3–4 % matched those in HES.

The concordance in diagnosis dates significantly improved over the years for both GOLD and Aurum when compared to CR and HES (Supplementary_information-S3). A higher proportion of diagnoses was recorded within a year when compared to CR (93 % in both GOLD and Aurum) versus HES (66 % in GOLD, 71 % in Aurum). Discrepancies were further explored by plotting the date differences only within a year, quantified in weeks, where positive x-axis values represented later dates in GOLD/Aurum and negative values earlier dates (Figs. 4 and 5). The distributions were similar for both GOLD and Aurum, with peaks at 2–3 weeks indicating delayed diagnoses in primary care compared to CR (Fig. 4). Most cases showed diagnoses recorded at least 3 months earlier in primary care compared to HES, with additional 2–3 week peaks (Fig. 5).

Table 3
Completeness estimates of prostate cancer diagnoses recorded in CPRD GOLD and CPRD Aurum compared to CR and HES.

Completeness estimates by patient characteristics compared to CR

Characteristics	CPRD GOLD			CPRD Aurum		
	Cases in CR (N)	Cases in GOLD and CR (N)	Completeness (%)	Cases in CR (N)	Cases in Aurum and CR (N)	Completeness (%)
Overall completeness estimates	23,804	15,443	64.9 %	21,123	16,172	76.6 %
Age at diagnosis						
< 60 years	680	561	82.5 %	928	848	91.4 %
60–69 years	7074	5200	73.5 %	6625	5650	85.3 %
70–79 years	10,266	6428	62.6 %	8654	6548	75.7 %
≥ 80 years	5784	3254	56.3 %	4916	3126	63.6 %
Year of diagnosis						
2000–2003	4681	3252	69.5 %	3984	2867	72.0 %
2004–2007	5364	4182	78.0 %	4766	3769	79.1 %
2008–2011	6078	4364	71.8 %	5323	4251	79.9 %
2012–2016	7681	3645	47.5 %	7050	5285	75.0 %
IMD						
1 - least deprived	3650	2349	64.4 %	4441	3420	77.0 %
2	3920	2518	64.2 %	4358	3412	78.3 %
3	5105	3387	66.3 %	4172	3174	76.1 %
4	5711	3597	63.0 %	4228	3216	76.1 %
5 - most deprived	5418	3592	66.3 %	3924	2950	75.2 %
Ethnicity						
Asian	182	110	60.4 %	277	216	78.0 %
Black	370	245	66.2 %	626	488	78.0 %
Mixed	49	35	71.4 %	74	64	86.5 %
Other	131	78	59.5 %	123	89	72.4 %
White	22,089	14,398	65.2 %	19,274	14,817	76.9 %
Practice Region						
1-North East	654	416	63.6 %	831	691	83.2 %
2-North West	1980	1382	69.8 %	3308	2595	78.4 %
3-Yorkshire & The Humber	1411	794	56.3 %	819	662	80.8 %
4-East Midlands	1101	578	52.5 %	416	304	73.1 %
5-West Midlands	1833	1259	68.7 %	4298	3374	78.5 %
6-East of England	3391	2050	60.5 %	1255	1002	79.8 %
7-South West	4876	3122	64.0 %	3409	2581	75.7 %
8-South Central	3346	2294	68.6 %	2541	1955	76.9 %
9-London	1884	1224	65.0 %	2267	1579	69.7 %
10-South East Coast	3328	2324	69.8 %	1927	1388	72.0 %

Completeness estimates by patient characteristics compared to HES

Characteristics	CPRD GOLD			CPRD Aurum		
	Cases in HES (N)	Cases in GOLD and HES (N)	Completeness (%)	Cases in HES (N)	Cases in Aurum and HES (N)	Completeness (%)
Overall completeness estimates	22,747	13,655	60.0 %	21,175	14,776	69.8 %
Age at diagnosis						
< 60 years	405	325	80.2 %	633	554	87.5 %
60–69 years	5236	3833	73.2 %	5366	4499	83.8 %
70–79 years	9177	5411	59.0 %	8100	5726	70.7 %
≥ 80 years	7929	4086	51.5 %	7076	3997	56.5 %
Year of diagnosis						
2000–2003	3818	2134	55.9 %	3598	2097	58.3 %
2004–2007	4777	3289	68.9 %	4503	3194	70.9 %
2008–2011	5973	4044	67.7 %	5603	4150	74.1 %
2012–2016	8179	4188	51.2 %	7471	5335	71.4 %
IMD						
1 - least deprived	3430	2058	60.0 %	4347	3029	69.7 %
2	3817	2242	58.7 %	4251	3041	71.5 %
3	4786	2951	61.7 %	4230	2955	69.9 %
4	5386	3165	58.8 %	4328	3004	69.4 %
5 - most deprived	5328	3239	60.8 %	4019	2747	68.4 %
Ethnicity						
Asian	182	100	54.9 %	279	197	70.6 %
Black	350	223	63.7 %	612	449	73.4 %
Mixed	38	27	71.1 %	67	51	76.1 %
Other	115	63	54.8 %	128	86	67.2 %
White	21,426	12,923	60.3 %	19,523	13,672	70.0 %
Practice Region						
1-North East	644	375	58.2 %	874	643	73.6 %
2-North West	2058	1345	65.4 %	3578	2588	72.3 %
3-Yorkshire & The Humber	1337	715	53.5 %	785	586	74.6 %
4-East Midlands	1069	533	49.9 %	395	279	70.6 %
5-West Midlands	1838	1143	62.2 %	4124	2960	71.8 %
6-East of England	3320	1881	56.7 %	1214	893	73.6 %
7-South West	4417	2680	60.7 %	3435	2466	71.8 %
8-South Central	2874	1793	62.4 %	2313	1576	68.1 %
9-London	1873	1098	58.6 %	2404	1484	61.7 %
10-South East Coast	3317	2092	63.1 %	2004	1271	63.4 %

Table 4
Discrepancies in prostate cancer diagnosis dates recorded in CPRD GOLD and CPRD Aurum compared to CR and HES.

	Comparison to CR		Comparison to HES	
	CPRD GOLD	CPRD Aurum	CPRD GOLD	CPRD Aurum
Median (IQR) (days)	15 (3–33)	15 (0–31)	−67 (−611–14)	−42 (−408–16)
Mean ± st. dev. (days)	12.3 ± 430	11.1 ± 447	−479.6 ± 984.8	−388.8 ± 901.9
Earlier in GOLD/Aurum	2614 (17 %)	3215 (20 %)	8492 (62 %)	8707 (59 %)
Same date	979 (6 %)	989 (6 %)	426 (3 %)	522 (4 %)
Later in GOLD/Aurum	11,850 (77 %)	11,968 (74 %)	4737 (35 %)	5547 (38 %)

3.6. Reasons diagnoses were missing in GOLD and Aurum

In the GOLD cohort (29,500 patients), 5939 (20.1 %) had prostate cancer (PCa) diagnoses recorded in both CR and HES but not in GOLD. Of them, 42 % had no records in GOLD within ±365 days of their CR diagnosis dates (Fig. 6). Around 7 % had supporting evidence such as treatment codes (e.g.: prostatectomy, radiotherapy etc.) indicating PCa management while 11 % had non-malignant PCa codes (e.g., in-situ, benign or uncertain behaviour). Another 4 % had no diagnosis codes but evidence of diagnostic tests for PCa (e.g.: PSA), and 6 % had other prostate conditions like BPH or prostatism. Among the remaining 4236 records, 1406 (24 %) had non-informative administrative codes such as 'attachment', 'referral letter' and 'scanned document', suggesting additional clinical details that likely contained diagnostic data but were inaccessible to researchers. For 6 % of cases, no relevant reason for missing diagnoses was identified.

In the Aurum cohort (26,475 patients), 3538 (13.4 %) were missing PCa diagnoses in Aurum but recorded in both CR and HES. Of them, 36 % had no Aurum records within ±365 days of their CR diagnosis date. Around 6 % had treatment codes indicating PCa care and 13 % had non-malignant PCa codes. Similar to GOLD, around 4 % had evidence of diagnostic tests without recorded diagnoses and 6 % had codes for other prostate conditions. Of the remaining 2522 records, 829 (23 %) had non-informative administrative codes, while 12 % had no identifiable reason for the missing diagnoses.

4. Discussion

This study provides a comprehensive evaluation of the quality of PCa diagnosis recording in primary care using CPRD GOLD and CPRD Aurum datasets by examining their accuracy, completeness, and recording dates using major national English data sources. We expected PCa diagnoses to be captured within linked external cancer registrations and secondary care data in addition to primary care, considering the integrated care pathway involving hospital treatment, cancer registry reporting, and follow-up care by GPs and specialists. Our findings reveal key differences in patient characteristics, data quality, and the value of data linkage for improving diagnosis capture while identifying challenges in

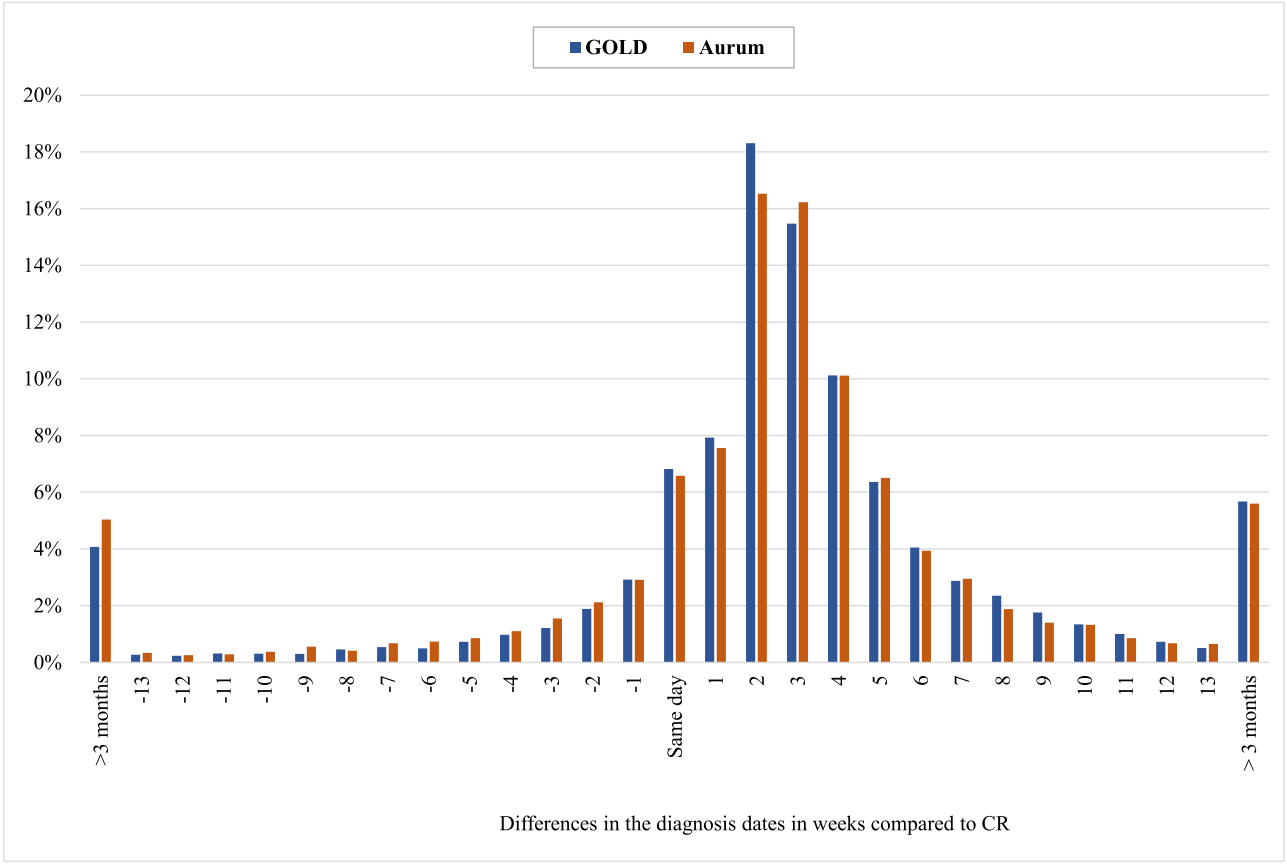


Fig. 4. Discrepancies in diagnosis dates recorded in GOLD and Aurum compared to CR.

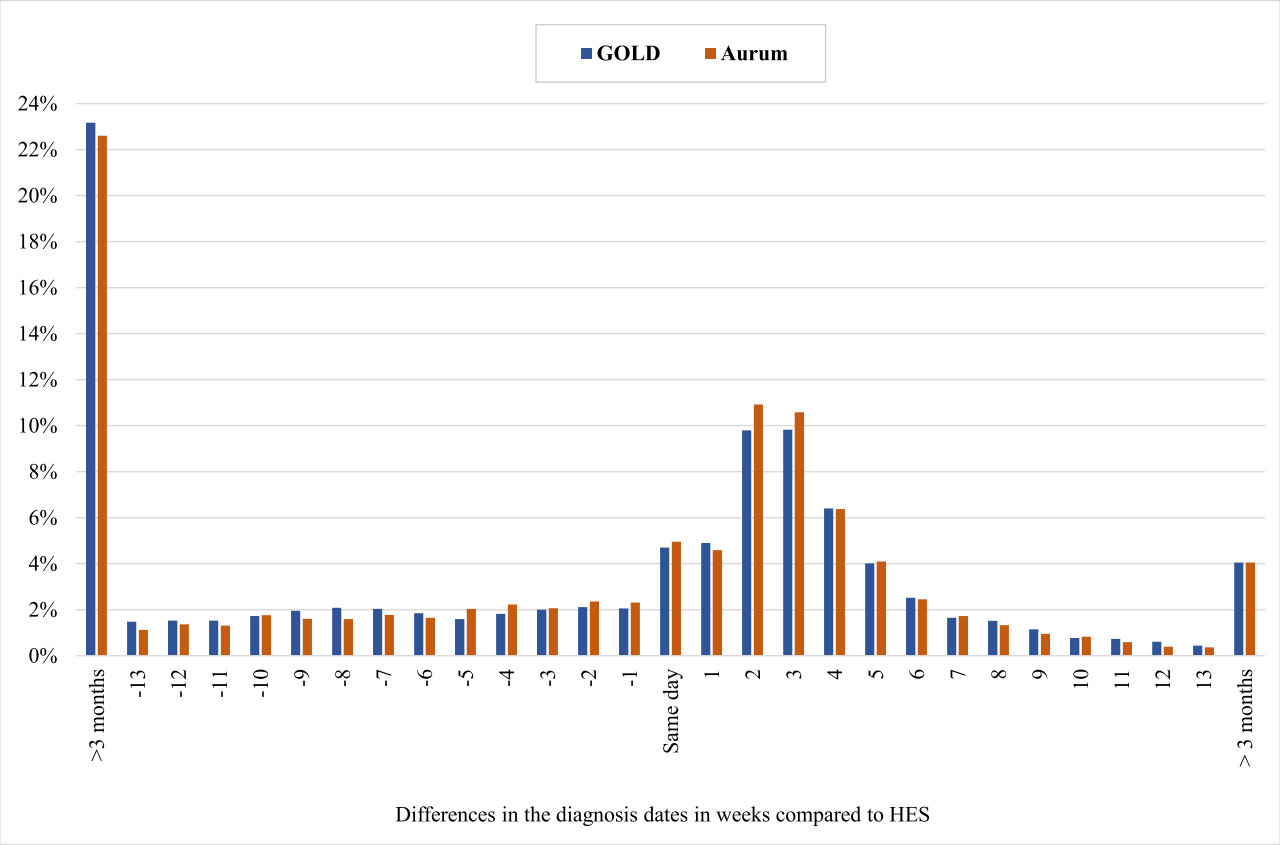


Fig. 5. Discrepancies in diagnosis dates recorded in GOLD and Aurum compared to HES.

data, particularly for older patient cohorts.

4.1. Comparison with previous studies

Our findings highlight that patient characteristics differ across data sources, with HES data capturing an older demographic (median age: 75–76 years). This aligns with previous studies suggesting that using HES alone may over-represent older patients [26,27]. The regional differences observed were also consistent with known geographic distributions of these datasets, with GOLD representing more patients from the South West, South Central, and South East Coast practices while Aurum covering more in the West Midlands, North West, and South West [1,18].

Both GOLD and Aurum showed higher accuracy and completeness estimates with CR compared to HES, consistent with prior findings [26,27]. However, our estimates were slightly lower than previously reported for GOLD [26–29] and Aurum [28,30], which may be attributed to our older cohorts, as increased age is known to be associated with discordance in cancer recording [28,30,31]. The overall accuracy and completeness, compared to combined CR and HES, were 93 % and 59 % for GOLD and 94 % and 69 % for Aurum. Margulis et al. reported similar accuracy but higher completeness (79 %) for GOLD, though their cohort which included only patients treated for overactive bladder, limits comparability [32].

Our study demonstrates the substantial benefits of data linkage for enhancing diagnosis capture. CR linkage could increase the capture of PCa diagnoses by 28 % in GOLD and 19 % in Aurum, while using HES could improve it by 31 % in GOLD and 24 % in Aurum. Combining both CR and HES could further increase diagnosis capture by 39 % in GOLD and 30 % in Aurum. However, around 9 % of cases identified in GOLD/Aurum had no corresponding records in CR and about 15 % were missing in HES. Moreover, we identified that around 20 % of cases recorded in either HES or the GOLD/Aurum record were not found in CR, a higher proportion than reported by Arhi et al. likely due to our older patient cohort [27]. Researchers often assume CR as a "gold standard" due to its importance in reporting cancer in England; however, its quality heavily varies by tumour type and calendar time [26,31]. Strongman et al. addressed this by developing a cancer algorithm without considering CR as an outright gold standard but used multiple data sources and identified cases in GOLD, HES, or ONS that were missing in CR [26]. Nevertheless, these missing cases in CR may also reflect provisional diagnoses, differences in definitions, diagnoses from private healthcare or outside England, or data linkage errors.

Concordance of diagnosis recording dates was also better with CR for both GOLD and Aurum (93 % within 1 year) than with HES (66–71 % within 1 year), improving over time, especially after 2004, likely due to initiatives like the Quality and Outcomes Framework (QOF) [33,34]. However, both GOLD and Aurum had most diagnoses recorded later than CR, with a median delay of 15 days which may be attributed to the time it takes for records to be processed in CR and then coded into the primary care system. HES comparison showed that most diagnoses occurred up to three months earlier or two to three weeks later in primary care. Arhi et al. observed a similar variation and suggested that the first peak, occurring over 3 months earlier in primary care may have reflected the extended period between neo-adjuvant treatment and surgery while delayed primary care diagnoses indicated potential diagnosis codes following resection, or an inpatient procedure in HES [27].

Among the missing primary care records, 52 % in GOLD and 45 % in Aurum were recorded in both CR and HES. Around 6–7 % of these records had evidence of treatment codes supporting PCa care, while 11–13 % had non-malignant diagnosis codes. Additionally, 23–24 % only included non-informative administrative codes, suggesting that although diagnostic information is communicated to primary care via

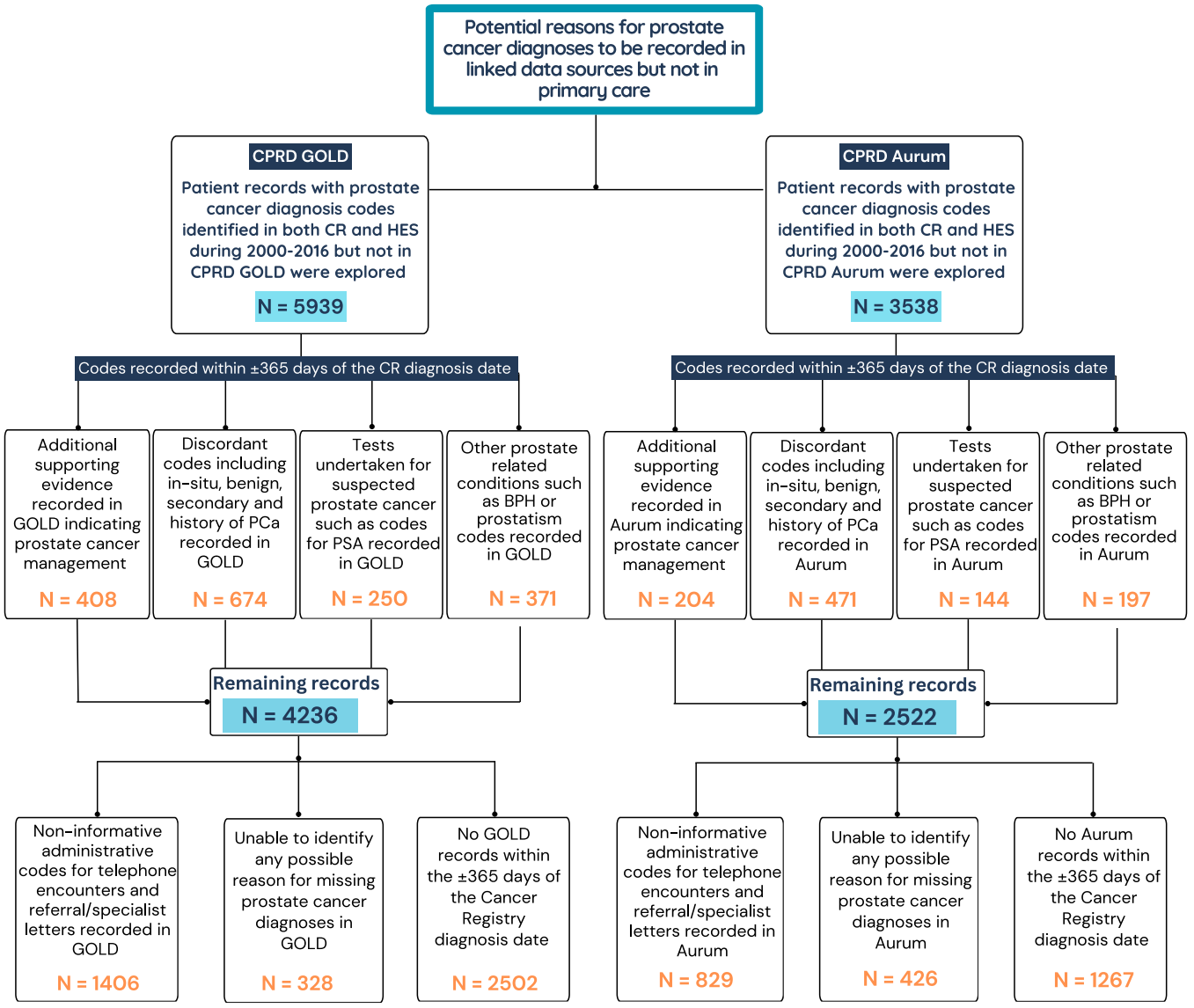


Fig. 6. Analysis of patient records where prostate cancer diagnoses were recorded in both CR and HES but not in GOLD and Aurum (within ±365 days of the CR diagnosis date).

letters and reports, it is often not properly coded. While clinicians can access these letters for clinical details on diagnoses, researchers face limitations due to the absence of coded data. This highlights the need for improving diagnostic coding in primary care or allowing researchers controlled access to clinical notes and letters to supplement the coded data to enhance data quality in research.

4.2. Strengths and limitations

To our knowledge, this is the first study to assess the quality of PCa diagnosis recording considering both GOLD and Aurum allowing cross-validation between the two databases. Previous studies have mostly focused on GOLD, and have not extensively compared Aurum with CR [28,30]. By estimating accuracy and completeness across patient demographics, we reported disparities in data quality in primary care, providing insights into the selection of datasets for future research. Investigating why patients had diagnostic codes in linked data but not in primary care highlighted challenges in case identification and the limitations of routinely collected data.

However, variations in study populations, linked data sources, timelines, prior cancer history and code definitions may have impacted our results. The inclusion of men aged 46 and above in the analysis may restrict the generalisability of the findings albeit the prevalence of PCa among men over 50 in the UK, with high age-specific incidence rates starting from 45 to 49 years and peaking in the 75–79 age group [13]. Moreover, the inclusion criteria for participants required them to be registered on their 65th birthday, which may have introduced selection bias in our study, as younger patients diagnosed with PCa would need to survive to be included. However, we believe the impact of this would be less given the high survival rates for PCa (84 % ten-year survival) [35].

The historical evolution of the data sources, particularly, CR and the changes made post-2013, may have influenced our findings. The applicability of the results might be further limited due to the study period which only extends to 2016, and the recording practices may have changed especially in light of the COVID-19 pandemic. Following the 2020 lockdown, cancer incidence experienced significant disruptions. Price et al. reported an estimated 14 % decrease in PCa incidence from March 2020 to February 2022, with levels consistently remaining below the expected throughout the entire period [36]. Given these changes, the insights drawn from our study may not fully represent the current landscape of PCa diagnosis recording, and ongoing evaluation of recording practices in primary care is necessary to ensure data quality

and accuracy in future research. Discrepancies in coding dictionaries are a potential limitation, although the use of predefined codes and team reviews with clinicians aimed to mitigate these issues. The exclusive use of PCa diagnostic codes without combining other domains (e.g., test/-morphology/treatment) may have affected case identification [30,37]. Reliance on structured data will miss PCa cases where diagnostic information exists in unstructured data in primary care clinical notes as we have reported.

While our study highlights the limitations in primary care data quality for PCa diagnoses, particularly among older men, it underscores the significant improvements achievable through data linkage. Our findings demonstrated higher recording accuracy and completeness in Aurum compared to GOLD across both CR and HES. Integrating multiple linked sources can greatly enhance the accuracy and completeness of PCa diagnoses, thereby strengthening the reliability of observational research. However, no single data source is a perfect reference standard and will capture all true cancer cases, hence researchers must account for the limitations of each source and the practical implications of linked data use. Data linkage may not always be suitable for every study, making quality assessment research, such as ours, valuable in understanding the strengths and limitations of each data source for more informed interpretation. Future research should further investigate PCa diagnosis recording quality in Aurum versus CR with a more representative cohort and explore reasons for unconfirmed or missing cases.

5. Conclusions

The study demonstrates better quality for PCa diagnosis recording in CPRD Aurum compared to CPRD GOLD affirming its suitability for research purposes. However, linking to external data sources such as CR or HES is recommended for complete case capture. Importantly, we observed low data quality within our study cohort of older patients, highlighting the need for careful consideration in studies focused on this demographic. Researchers must address inherent limitations within each data source, tailor approaches to study requirements and mitigate disparities in recording, timing, and patient characteristics to ensure research validity.

Authors' contributions

GS, AL, and EF contributed to the conceptualisation of the study; GS carried out the main analysis and wrote the manuscript; AL, EF, JA and SM provided advice on the methodology and completed the subsequent revisions and editing of the manuscript; MC and PF provided advice on the data. All authors commented on successive drafts, and read and approved the final manuscript.

Authorship contribution statement

Gayasha Somathilake (GS), Agnieszka Lemanska (AL), and Elizabeth Ford (EF) contributed to the conceptualisation of the study; GS carried out the main analysis and wrote the revised manuscript; AL, EF, Jo Armes (JA) and Sotiris Moschoyiannis (SM) provided advice on the methodology and completed the subsequent revisions and editing of the manuscript; Michelle Collins (MC) and Patrick Francsics (PF) provided advice on the data. All authors commented on successive drafts and read and approved the final manuscript.

CRediT authorship contribution statement

Gayasha Batheegama Gamarachchige: Writing – original draft, Writing – review & editing, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Elizabeth Ford:** Writing – review & editing, Validation, Supervision, Software, Resources, Methodology, Conceptualization. **Jo Armes:** Writing – review & editing, Validation, Supervision. **Sotiris Moschoyiannis:** Writing – review & editing, Validation, Supervision. **Michelle Collins:** Writing – review & editing, Methodology. **Patrick Francsics:** Writing – review & editing, Methodology. **Agnieszka Lemanska:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Footnotes

None

Provenance and peer review

Not commissioned, externally peer-reviewed.

Additional Information

Supplementary material available in Supplementary Information *(SI).pdf*

Correspondence and requests for materials should be addressed to Gayasha Somathilake. Software developed for data extraction and the code lists used in this study will be made openly accessible for review and reuse on GitHub (https://github.com/).

Declaration of Competing Interest

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.

Acknowledgements

This research was funded by the University of Surrey, Doctoral College - UK as part of the Breaking Barriers Doctoral studentship awarded to GS. The work of NPL co-authors was partly funded by the UK Government's Department for Science, Innovation & Technology through the UK's National Measurement System programmes via the Semantic Technologies theme. JA, EF amd SM receive funding from the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Kent, Surrey, Sussex. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

This study is based on data from the CPRD obtained under licence from the UK Medicines and Healthcare Products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The study was approved by the Independent Scientific Advisory Committee for CPRD research (protocol 19_050R). The interpretation and conclusions contained in this study are those of the authors alone.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.canep.2024.102715.

References

[1] E. Herrett, A.M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa, et al., Data resource profile: clinical practice research datalink (CPRD), Int. J. Epidemiol. [Internet] (2015). ⟨https://academic.oup.com/ije/article/44/3/827/632531⟩ [cited 2022 Dec 14];44(3):827–36. Available from.

[2] Vezyridis P., Timmons S. Evolution of primary care databases in UK: a scientometric analysis of research output. [cited 2024 Jan 22]; Available from: https://doi.org/10.1136/bmjopen-2016-012785.

[3] L. Wang, S. Fu, A. Wen, Xiaoyang Ruan, H. He, S. Liu, et al., Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing, JCO Clin. Cancer Inf. [Internet]. 2022 6 (2024) 2200006. Available from: https://doi.org/10.

[4] Cowie M.R., Blomster J.I., Curtis L.H., Duclaux S., Ford I., Fritz F., et al. Electronic health records to facilitate clinical research. Clinical Research in Cardiology

[Internet]. 2017 Jan 1 [cited 2024 Jan 22];106(1):1–9. Available from: ⟨https://link.springer.com/article/10.1007/s00392-016-1025-6⟩.
[5] Hamilton W. Cancer diagnosis in primary care. British Journal of General Practice [Internet]. 2010 Feb 1 [cited 2024 Jan 24];60(571):121–8. Available from: ⟨https://bjgp.org/content/60/571/121⟩.
[6] Watt T., Sullivan R., Aggarwal A. Primary care and cancer: an analysis of the impact and inequalities of the COVID-19 pandemic on patient pathways. [cited 2024 Jan 24]; Available from: https://doi.org/10.1136/bmjopen-2021-059374.
[7] Tayefi M., Ngo P., Chomutare T., Dalianis H., Salvi E., Budrionis A., et al. Challenges and opportunities beyond structured data in analysis of electronic health records. 2021 [cited 2024 Jan 24]; Available from: https://doi.org/10.1002/wics.1549.
[8] B.A. Goldstein, A.M. Navar, M.J. Pencina, J.P.A. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. [Internet] (2017). ⟨https://academic.oup.com/jamia/article/24/1/198/2631444⟩ [cited 2023 Jan 22];24(1):198–208. Available from.
[9] CPRD bibliography. [cited 2024 Jan 26]; Available from: https://doi.org/10.1007/s10151-023-02874-3. .
[10] Wing K., Bhaskaran K., Smeeth L., Van Staa T.P., Klungel O.H., Reynolds R.F., et al. Optimising case detection within UK electronic health records: use of multiple linked databases for detecting liver injury. Available from: ⟨http://bmjopen.bmj.com/⟩.
[11] E. Badrick, I. Renehan, A.G. Renehan, Linkage of the UK Clinical Practice Research Datalink with the national cancer registry, Eur. J. Epidemiol. [Internet] (2019). ⟨https://link.springer.com/article/10.1007/s10654-018-0441-5⟩ [cited 2024 Jul 19];34(1):101–2. Available from.
[12] E. Herrett, S.L. Thomas, W. Marieke Schoonen, L. Smeeth, A.J. Hall, M. Emily Herrett, 2010, Validation and validity of diagnoses in the General Practice Research Database: a systematic review. [cited 2024 Jan 29]; Available from: ⟨http://www3.interscience.wiley.com/⟩⟨http://www3.interscience.wiley.com/⟩.
[13] Prostate cancer | Cancer Research UK [Internet]. [cited 2024 Jan 27]. Available from: ⟨https://www.cancerresearchuk.org/about-cancer/prostate-cancer⟩.
[14] Merriel S.W.D., Seggie A., Ahmed H. Diagnosis of prostate cancer in primary care: navigating updated clinical guidance. British Journal of General Practice [Internet]. 2023 Feb 1 [cited 2024 Jan 27];73(727):54–5. Available from: ⟨https://bjgp.org/content/73/727/54⟩.
[15] Nicholson A., Ford E., Davies K.A., Smith H.E., Rait G., Tate A.R., et al. Optimising Use of Electronic Health Records to Describe the Presentation of Rheumatoid Arthritis in Primary Care: A Strategy for Developing Code Lists. [cited 2024 Mar 16]; Available from: ⟨www.plosone.org⟩.
[16] K.E. Henson, L. Elliss-Brookes, V.H. Coupland, E. Payne, S. Vernon, B. Rous, et al., Data resource profile: national cancer registration dataset in England, Int. J. Epidemiol. [Internet] (2020). ⟨https://academic.oup.com/ije/article/49/1/16/5476570⟩ [cited 2023 Jan 5];49(1):16–16h. Available from.
[17] A. Herbert, L. Wijlaars, A. Zylbersztejn, D. Cromwell, P. Hardelid, Data resource profile: hospital episode statistics admitted patient care (HES APC), Int. J. Epidemiol. [Internet] (2017). ⟨https://pubmed.ncbi.nlm.nih.gov/28338941/⟩ [cited 2023 Jan 5];46(4):1093–1093i. Available from.
[18] A. Wolf, D. Dedman, J. Campbell, H. Booth, D. Lunn, J. Chapman, et al., Data resource profile: clinical practice research datalink (CPRD) aurum, Int. J. Epidemiol. [Internet] (2019). ⟨https://academic.oup.com/ije/article/48/6/1740/5374844⟩ [cited 2023 Jan 5];48(6):1740–1740g. Available from.
[19] D. Dedman, H. Strongman, S. Hodgson, R.E. Ghosh, 2019, Small area level data based on practice postcode Documentation.
[20] SNOMED CT Starter Guide - SNOMED CT Starter Guide - SNOMED Confluence [Internet]. [cited 2024 Oct 9]. Available from: ⟨https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide⟩.
[21] Home | SNOMED International [Internet]. [cited 2024 Oct 9]. Available from: ⟨https://www.snomed.org/⟩.
[22] Read Codes - NHS Digital [Internet]. [cited 2023 Jan 5]. Available from: ⟨https://digital.nhs.uk/services/terminology-and-classifications/read-codes⟩.
[23] Chisholm J. The Read clinical classification. BMJ [Internet]. 1990 [cited 2024 Oct 9];300(6732):1092. Available from: ⟨https://pubmed.ncbi.nlm.nih.gov/2344534/⟩.
[24] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. [Internet] (2013). ⟨https://academic.oup.com/jamia/article/20/1/144/2909176⟩ [cited 2023 Jan 31];20(1):144–51. Available from.
[25] Plos Medicine |, Doi ; Benchimol E.I., Smeeth L., Guttmann A., Harron K., Moher D., et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med [Internet]. 2015 [cited 2024 Jan 21];6(10):1001885. Available from: ⟨http://www.record-statement.org⟩.
[26] H. Strongman, R. Williams, K. Bhaskaran, What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? a concordance and validation study using linked English electronic health records data, BMJ Open 10 (8) (2020).
[27] C.S. Arhi, A. Bottle, E.M. Burns, J.M. Clarke, P. Aylin, P. Ziprin, et al., Comparison of cancer diagnosis recording between the Clinical Practice Research Datalink, Cancer Registry and Hospital Episodes Statistics, Cancer Epidemiol. 57 (2018 Dec 1) 148–157.
[28] A.M. Trafford, R. Parisi, M.K. Rutter, E. Kontopantelis, C.E.M. Griffiths, D. M. Ashcroft, Concordance and timing in recording cancer events in primary care, hospital and mortality records for patients with and without psoriasis: a population-based cohort study, PLoS One 16 (7 July 2021) (2021 Jul 1).
[29] R. Williams, T.P. Van Staa, A.M. Gallagher, T. Hammad, H.G.M. Leufkens, F. De Vries, 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 8 (5) (2018).
[30] K.W. Hagberg, C. Vasilakis-Scaramozza, R. Persson, E. Yelland, T. Williams, P. Myles, et al., Quality and completeness of malignant cancer recording in United Kingdom Clinical Practice Research Datalink Aurum compared to Hospital Episode Statistics, Ann. Cancer Epidemiol. 6 (2022), 6–6.
[31] K.W. Hagberg, C. Vasilakis-Scaramozza, R. Persson, D. Neasham, G. Kafatos, S. Jick, Correctness and completeness of breast cancer diagnoses recorded in UK CPRD aurum and CPRD GOLD databases: comparison to hospital episode statistics and cancer registry (Companion Paper 2), Clin. Epidemiol. 15 (2023) 1193–1206.
[32] A.V. Margulis, J. Fortuny, J.A. Kaye, B. Calingaert, M. Reynolds, E. Plana, et al., Validation of cancer cases using primary care, cancer registry, and hospitalization data in the United Kingdom, Epidemiol. [Internet] (2018) [cited 2024 Jan 26];29(2):308. Available from: /pmc/articles/PMC5794229/.
[33] QOF guidance for 2023/24. [cited 2024 Jan 30]; Available from: ⟨https://www.gov.uk/government/publications/nhs-primary-medical-services-directions-2013⟩.
[34] Taggar J.S., Coleman T., Lewis S., Szatkowski L. The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: cross-sectional analyses from The Health Improvement Network (THIN) database. 2012 [cited 2024 Oct 2]; Available from: ⟨http://www.biomedcentral.com/1471-2458/12/329⟩.
[35] M. Quaresma, M.P. Coleman, B. Rachet, 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study, Lancet [Internet] (2015 Mar 28). ⟨https://pubmed.ncbi.nlm.nih.gov/25479696/⟩ [cited 2024 Oct 10];385(9974):1206–18. Available from.
[36] S. Price, S. Bailey, W. Hamilton, D. Jones, L. Mounce, G. Abel, The effects of the first UK lockdown for the COVID-19 pandemic on primary-care-recorded cancer and type-2 diabetes mellitus records: a population-based quasi-experimental time series study, Cancer Epidemiol. 91 (2024 Aug 1) 102605.
[37] Nicholson A., Ford E., Davies K.A., Smith H.E., Rait G., Tate A.R., et al. Optimising Use of Electronic Health Records to Describe the Presentation of Rheumatoid Arthritis in Primary Care: A Strategy for Developing Code Lists. [cited 2024 Mar 20]; Available from: ⟨www.plosone.org⟩.