Original Article

Quality and completeness of malignant cancer recording in United Kingdom Clinical Practice Research Datalink Aurum compared to Hospital Episode Statistics

Katrina Wilcox Hagberg[1], Catherine Vasilakis-Scaramozza[1], Rebecca Persson[1], Eleanor Yelland[2], Tim Williams[2], Puja Myles[2], Susan S. Jick[1,3]^

[1]Boston Collaborative Drug Surveillance Program, Lexington, MA, USA; [2]Clinical Practice Research Datalink, Medicines and Healthcare products Regulatory Agency, London, UK; [3]Boston University School of Public Health, Boston, MA, USA

Contributions: (I) Conception and design: All authors; (II) Administrative support: P Myles, SS Jick; (III) Provision of study materials or patients: E Yelland, T Williams, P Myles; (IV) Collection and assembly of data: E Yelland, T Williams, P Myles; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Susan S. Jick, ScD. Boston Collaborative Drug Surveillance Program, 11 Muzzey St, Lexington, MA 02421, USA. Email: sjick@bu.edu.

Background: Data source validation is necessary to inform suitability for use in medical research. The objective was to examine agreement of cancer diagnoses recorded in Clinical Practice Research Datalink (CPRD) Aurum compared with linked Hospital Episode Statistics (HES) data to provide information on CPRD Aurum data correctness (accuracy, validity) and completeness (presence, missingness).

Methods: The source population was a 50,000 random sample of CPRD Aurum patients with HES linkage (1997–2017). Patients with cancer diagnoses recorded in either data source were selected for these analyses. Correctness was the proportion of patients with cancer recorded in CPRD Aurum with a concordant cancer diagnosis recorded in HES. Completeness was the proportion of patients with cancer recorded in HES with a concordant diagnosis in CPRD Aurum.

Results: A total of 6,019 patients had a cancer diagnosis: 3,864 in CPRD Aurum, 5,545 in HES, and 3,390 in both. Correctness estimate was 87.7% and an additional 8.4% had supporting cancer codes recorded in CPRD Aurum. Completeness was 61.1% and an additional 22.7% had supporting cancer codes recorded in CPRD Aurum. Correctness and completeness estimates varied by cancer site and calendar time.

Conclusions: Cancer diagnoses in CPRD Aurum were of relatively high correctness for use in medical research. Completeness varied by cancer type and calendar year and may not be sufficient for all research questions. Use of linked data may improve completeness.

Keywords: Clinical Practice Research Datalink (CPRD); CPRD Aurum; data completeness; data quality; cancer

Received: 12 March 2022; Accepted: 11 August 2022; Published: 30 September2022.

doi: 10.21037/ace-22-4

View this article at: https://dx.doi.org/10.21037/ace-22-4

Introduction

The United Kingdom (UK) Clinical Practice Research Datalink (CPRD) Aurum is an electronic primary care database sourced from Egton Medical Information Systems (EMIS®) patient management software which became available in 2018 (1). While there are similarities between CPRD Aurum and CPRD GOLD, another primary care data source with well-established reliability and quality for use in medical research, the quality of recording in CPRD Aurum has yet to be fully assessed (2-8). Assessments of the quality and completeness of all new data resources are necessary to evaluate suitability for medical research.

^ ORCID: 0000-0002-2215-1067.

We have previously published validation assessments describing recording of pulmonary embolism, myocardial infarction, and diabetes, hypercholesterolemia, and anemia in CPRD Aurum using methodologies described by Weiskopf and Weng (9-12). This study used the same patient population as prior assessments to describe data source agreement on the presence of malignant cancer diagnoses recorded in CPRD Aurum primary care data compared with Hospital Episode Statistics (HES) Admitted Patient Care (APC) data, which has the most complete capture of diagnoses and procedures provided in the hospital settings (13,14). This comparison provides information on "correctness" (i.e., accuracy, validity) and "completeness" (i.e., presence, missingness) of cancer diagnoses recorded in CPRD Aurum. For most cancers, we expect diagnoses to appear in both data sources because HES APC data captures diagnoses and procedures conducted in-hospital and follow-up care is provided by general practitioners (GPs) (13,14). This study provides an assessment of the quality and completeness of cancer diagnoses recorded in CPRD Aurum, but the results may also be an indicator of the quality of recording of other chronic conditions with similar clinical care pathways. We present the following article in accordance with the STROBE reporting checklist (available at https://ace.amegroups.com/article/view/10.21037/ace-22-4/rc).

Methods

Data resources

CPRD Aurum is provided by CPRD, a research service jointly supported by the Medicines and Healthcare products Regulatory Agency and the National Institute for Health Research, as part of the UK Department of Health and Social Care. As described in prior publications, CPRD Aurum is a large, prospectively collected, population-based, anonymized electronic medical record database (1,9-11). GPs record demographic information, prescription details, clinical events, referrals, hospital admissions, laboratory results, and lifestyle details (e.g., smoking, alcohol consumption) using EMIS® patient management software. As gatekeepers for all National Health Service (NHS) care, including hospital and specialist referrals, GP records are expected to include primary diagnoses leading to hospital referrals and details of encounters at secondary care providers (8). Data for this study was extracted in November 2018.

HES APC data was used as an external reference standard for this validation study. HES APC contains information on inpatient hospitalizations in England since 1997 for the purpose of hospital payment (13,14). CPRD Aurum practices in England are linked to HES APC data. HES APC data contains details of each NHS hospital stay, including diagnoses made during the stay, procedures performed, and dates of admission and discharge.

Study population

The source population was a random sample of 50,000 CPRD Aurum patients from among practices with a recent HES APC update in October 2018. To enable comparison of data recordings, patients in the source population were required to have at least one admission for any reason recorded in HES APC after the latest of the following: patient's last EMIS registration date, the patient's 20th birthday based on year of birth, or the start of HES coverage (April 1, 1997). This 50,000-patient sample was also used for other CPRD Aurum validation studies that describe other data elements and outcomes (9-11).

The study period was April 1, 1997, through December 31, 2017 (time frame when data from both sources was present). The start and end of each patient's active CPRD Aurum electronic record were estimated using available registration, prescription, and clinical data [Supplementary file (Appendix 1): Start End]. Patient's cohort entry date was defined as April 1, 1997 (start of HES data) or their estimated CPRD Aurum record start date, whichever came later. The end of follow-up was defined as first of the patient's estimated CPRD Aurum end date, death date, or December 31, 2017 (end of HES data). We excluded patients whose CPRD Aurum and HES APC record did not overlap or who did not have a valid birth date. We also excluded patients with a record of a prior cancer diagnosis in either data source before cohort entry because recording of cancer may vary based on prior cancer history in either data source.

Cancer diagnosis identification

To align coding systems between the two data sources, CPRD Aurum MedCodes were organized to match ICD-10 neoplasm groupings at specified cancer sites (available online: https://cdn.amegroups.cn/static/public/ace-22-4-1.xlsx: codes) (15). We did not evaluate cancers at ill-defined and unspecified sites, *in situ* or benign neoplasms,

or neoplasms with unspecified behavior. We selected all patients with a first-time code for cancer at a specified site recorded in either CPRD Aurum or HES APC after cohort entry.

Statistical analyses

We assessed "correctness" of cancer diagnoses in CPRD Aurum as the proportion of patients with at least one cancer diagnosis at a specified site in CPRD Aurum that also had a concordant diagnosis recorded in HES APC, the external reference standard (12). We report correctness overall, by cancer site, and stratified by sex. We also described the timing of cancer diagnosis coding between CPRD Aurum and HES APC. We then restricted the assessment to patients who in addition to a cancer diagnosis code in CPRD Aurum also had supporting clinical codes related to cancer care, chemotherapy, radiology, referrals/specialist visits, and palliative care.

To assess "completeness" of cancer diagnoses recorded in CPRD Aurum, we calculated the proportion of patients who had a concordant diagnosis present in CPRD Aurum (12). We report completeness overall, by cancer site, and stratified by sex.

For both correctness and completeness, we reviewed the electronic records for patients who had a diagnosis of cancer coded in only one of the two data sources and described potential explanations for differences in recording, including issues with data integrity and presence of other supporting clinical codes.

All analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA).

Ethical review and copyright

This study is based in part on data from the Clinical Practice Research Datalink obtained under license 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 interpretation and conclusions contained in this study are those of the authors alone. This study was approved by the Independent Scientific Advisory Committee (ISAC) for Medicines and Healthcare Products Regulatory Agency (protocol No: 18_191), and the protocol was made available to the journal reviewers upon request. This study used anonymized electronic medical records, no patient contact occurred in its conduct, and it was performed in accordance with the Declaration of Helsinki (as revised in 2013). Hospital Episode Statistics (HES) Copyright© (2018), re-used with the permission of The Health & Social Care Information Centre. All rights reserved. Researchers can apply for a limited license to access CPRD data for public health research, subject to individual research protocols meeting CPRD data governance requirements. More details including data specification, license fees and applications process are available on the CPRD website (https://www.cprd.com).

Results

Study population and characteristics of cancer patients

From the 50,000-patient source population, we excluded 581 (1.2%) patients whose CPRD Aurum and HES APC record did not overlap and <0.1% (N<5, not reportable) patients who did not have a valid birth date. We also excluded 1,704 (3.4%) patients with a prior cancer diagnosis recorded in either data source before cohort entry. From among the remaining 47,771 eligible patients, there were 6,019 (12.6%) patients with a diagnosis code for cancer at a specified site: 3,864 had a diagnosis coded in CPRD Aurum, 5,545 had a diagnosis in HES APC, and 3,390 had a cancer diagnosis code in both data sources. Patient sex, year of first cancer diagnosis, age at first cancer diagnosis, and follow-up time were similar for patients with a cancer record in CPRD Aurum and/or HES APC (*Table 1*).

Correctness of cancer diagnoses recorded in CPRD Aurum

There were 3,864 patients who had a code for cancer at a specified site recorded in CPRD Aurum, of which 3,390 (87.7%) also had a concordant diagnosis at the same site recorded in HES APC. Correctness was greater than 80% regardless of diagnosis year and age at first cancer diagnosis (*Table 2*). The cancer diagnosis date recorded in CPRD Aurum corresponded closely with the diagnosis date recorded in HES APC (median difference 24 days, interquartile range 10–82 days): 265 (6.9%) had the diagnosis recorded on the same date, 1,664 (43.1%) recorded 1–30 days apart, 651 (16.9%) 31–90 days apart, and 1,284 (33.2%) were recorded more than 90 days apart. When we assessed correctness by cancer site, the proportion of patients who had a diagnosis code for that site recorded in both CPRD Aurum and HES APC remained greater than 80% for most cancer sites. Correctness was highest

Table 1 Characteristics of patients with a first-time cancer at a specified cancer site in 50,000 CPRD Aurum patient sample, by data source

Characteristic	Cancer cases in CPRD Aurum sample (N=3,864) (%)	Cancer cases in HES APC[†] (N=5,545) (%)
Sex		
Female	1,868 (48.3)	2,635 (47.5)
Male	1,996 (51.7)	2,910 (52.5)
Year of first cancer diagnosis		
1997–1999	237 (6.1)	500 (9.0)
2000–2004	831 (21.5)	1,144 (20.6)
2005–2009	980 (25.4)	1,373 (24.8)
2010–2014	1,113 (28.8)	1,542 (27.8)
2015–2017	703 (18.2)	986 (17.8)
Age at first cancer diagnosis (years)		
20–29	32 (0.8)	41 (0.7)
30–39	116 (3.0)	150 (2.7)
40–49	258 (6.7)	330 (6.0)
50–59	621 (16.1)	800 (14.4)
60–69	1,003 (26.0)	1,347 (24.3)
70–79	1,065 (27.6)	1,539 (27.8)
≥80	769 (19.9)	1,338 (24.1)
Follow-up time[‡] (years)		
Mean ± St. Dev.	13.1±6.3	12.6±6.6
Median	13.7	13.0
Interquartile range	7.7–19.8	6.8–19.4

[†], HES APC matched to the CPRD Aurum 50,000 patient sample; [‡], patients followed from 1 April 1997 (start of HES APC data) or the start of the patient's electronic record (whichever came later) through 12/31/2017 12 December 2017 (end of HES APC data) or the end of the patient's electronic record (whichever came first). CPRD, Clinical Practice Research Datalink; HES APC, Hospital Episode Statistics Admitted Patient Care; St. Dev., standard deviation.

for cancers of digestive organs (92.2%), cancers of lip, oral cavity, and pharynx (85.9%), respiratory and intrathoracic organs (85.5%), urinary tract (85.2%), breast (83.7%), and cancers of male genital organs (80.9%). Correctness was lowest for thyroid and other endocrine glands (66.7%), melanoma and other malignant neoplasms of the skin (70.4%), and cancers of bone and articular cartilage (74.4%) (*Table 2*). Overall, correctness was slightly higher for males (89.3%) than females (86.1%). Correctness was higher for males for cancers of lip, oral cavity, and pharynx (91.5% versus 75.0% females), bone and articular cartilage (73.9% versus 68.5%), and urinary tract (89.7% versus 73.4%) (*Table 2*).

Approximately 85% of patients with cancer diagnoses at a specific site recorded in CPRD Aurum also had other supporting clinical codes consistent with cancer diagnosis or care in their CPRD Aurum record (e.g., suspected cancer codes, cancer diagnosis, cancer care, chemotherapy, referrals, specialist visits, palliative care) that supported the presence of cancer ("true cases") (*Table 2*). When we restricted the cases in CPRD Aurum to those who had supporting clinical codes, 88.6% had a concordant diagnosis recorded in HES APC. While this correctness estimate for all cancers at a specified site (88.6%) was similar to that found in the main analysis (87.7%), correctness estimates were improved for some less common cancer sites (i.e., bone and articular cartilage and thyroid or other endocrine glands) when we restricted the assessment to patients who

Table 2 Recording of cancer diagnoses at a specified site in a sample of 50,000 patients from CPRD Aurum, compared to HES APC (correctness), overall, stratified by sex, and restricted to patients with supporting clinical codes

Category	All patients with cancer code at a specified site in CPRD Aurum 50,000 patient sample			Female patients with cancer code at a specified site in CPRD Aurum 50,000 patient sample			Male patients with cancer code at a specified site in CPRD Aurum 50,000 patient sample			Restricted to patients with cancer code at a specified site plus supporting clinical codes[†] in CPRD Aurum 50,000 patient sample		
	N in CPRD Aurum sample	N in CPRD Aurum and HES APC	Correctness (%)	N in CPRD Aurum sample	N in CPRD Aurum and HES APC	Correctness (%)	N in CPRD Aurum sample	N in CPRD Aurum and HES APC	Correctness (%)	N in CPRD Aurum sample	N in CPRD Aurum and HES APC	Correctness (%)
Overall[‡]	3,864	3,390	87.7	1,868	1,608	86.1	1,996	1,782	89.3	3,302	2,925	88.6
By year of first cancer diagnosis												
1997–1999	237	197	83.1	130	104	80.0	107	93	86.9	115	93	80.9
2000–2004	831	720	86.6	403	339	84.1	428	381	89.0	568	505	88.9
2005–2009	980	878	89.6	445	387	87.0	535	491	91.8	901	809	89.8
2010–2014	1,113	995	89.4	537	482	89.8	576	513	89.1	1,049	943	89.9
2015–2017	703	600	85.4	353	296	83.9	350	304	86.9	669	575	86.0
By age at first cancer diagnosis (years)												
20–49	406	338	83.3	278	229	82.4	128	109	85.2	358	304	84.9
50–59	621	552	88.9	365	322	88.2	256	230	89.8	547	490	89.6
60–69	1,003	896	89.3	457	398	87.1	546	498	91.2	888	795	89.5
70–79	1,065	939	88.2	409	354	86.6	656	585	89.2	906	803	88.6
≥80	769	665	86.5	359	305	85.0	410	360	87.8	603	533	88.4
Malignant neoplasms, by site[§]												
Lip, oral cavity and pharynx	71	61	85.9	24	18	75.0	47	43	91.5	61	55	90.2
Digestive organs	977	901	92.2	399	355	89.0	578	546	94.5	807	745	92.3
Colon	338	269	79.6	146	119	81.5	192	150	78.1	291	230	79.0
Liver	201	172	85.6	87	69	79.3	114	103	90.4	178	153	85.5
Pancreas	97	85	87.6	33	28	84.9	64	57	89.1	70	62	88.6
Rectum and Anus	231	185	80.1	95	74	77.9	136	111	81.6	194	158	81.4
Stomach, esophagus and small intestines	217	199	91.7	69	61	88.4	148	138	93.2	169	156	92.3
Respiratory and intrathoracic organs	530	453	85.5	219	189	86.3	311	264	84.9	432	374	86.6

Table 2 (*continued*)

Table 2 (*continued*)

Category	All patients with cancer code at a specified site in CPRD Aurum 50,000 patient sample			Female patients with cancer code at a specified site in CPRD Aurum 50,000 patient sample			Male patients with cancer code at a specified site in CPRD Aurum 50,000 patient sample			Restricted to patients with cancer code at a specified site plus supporting clinical codes[†] in CPRD Aurum 50,000 patient sample		
	N in CPRD Aurum sample	N in CPRD Aurum and HES APC	Correctness (%)	N in CPRD Aurum sample	N in CPRD Aurum and HES APC	Correctness (%)	N in CPRD Aurum sample	N in CPRD Aurum and HES APC	Correctness (%)	N in CPRD Aurum sample	N in CPRD Aurum and HES APC	Correctness (%)
Lung	507	422	83.2	218	181	83.0	289	241	83.4	412	347	84.2
Other respiratory and intrathoracic organs	28	19	67.9	NR	NR	50	26	18	69.2	25	17	68.0
Bone and articular cartilage	43	32	74.4	20	15	68.5	23	17	73.9	37	31	83.8
Melanoma and other malignant neoplasms of skin	216	152	70.4	111	76	68.5	105	76	72.4	175	124	70.9
Mesothelial and soft tissue	39	31	79.5	15	11	83.7	24	20	83.3	33	27	81.8
Breast	743	622	83.7	738	618	83.7	5	NR	80	696	585	84.1
Female genital	219	182	83.1	217	182	83.9	n/a	n/a	n/a	197	168	85.3
Male genital organs	674	545	80.9	n/a	n/a	n/a	674	545	80.9	619	505	81.6
Prostate	639	518	81.1	n/a	n/a	n/a	639	518	80.9	587	480	81.8
Other	283	26	72.2	n/a	n/a	n/a	36	26	72.2	33	24	72.7
Urinary tract	283	241	85.2	79	58	73.4	204	183	89.7	240	206	85.8
Eye, brain and other parts of central nervous system	133	102	76.7	67	51	76.1	66	51	77.3	111	87	78.4
Thyroid and other endocrine glands	21	18	66.7	14	11	78.6	7	7	100	20	18	90.0
Lymphoid, hematopoietic and related tissues	324	259	79.9	146	120	82.2	178	139	78.1	258	207	80.2

[†], supporting clinical code defined as having one or more codes related to cancer care, chemotherapy, radiology, referrals/specialist visits, and palliative care referrals, specialist visits, or palliative care recorded in CPRD Aurum; [‡], cancer was required to be at the same specified site. Metastatic, *in situ* and benign and neoplasms, ill-defined and unspecified sites, and unspecified behavior were not evaluated in this study; [§], patients may have cancer at more than one site. CPRD, Clinical Practice Research Datalink; HES APC, Hospital Episode Statistics Admitted Patient Care; NR, not reported due to cell counts of 1–4 people; n/a, not applicable.

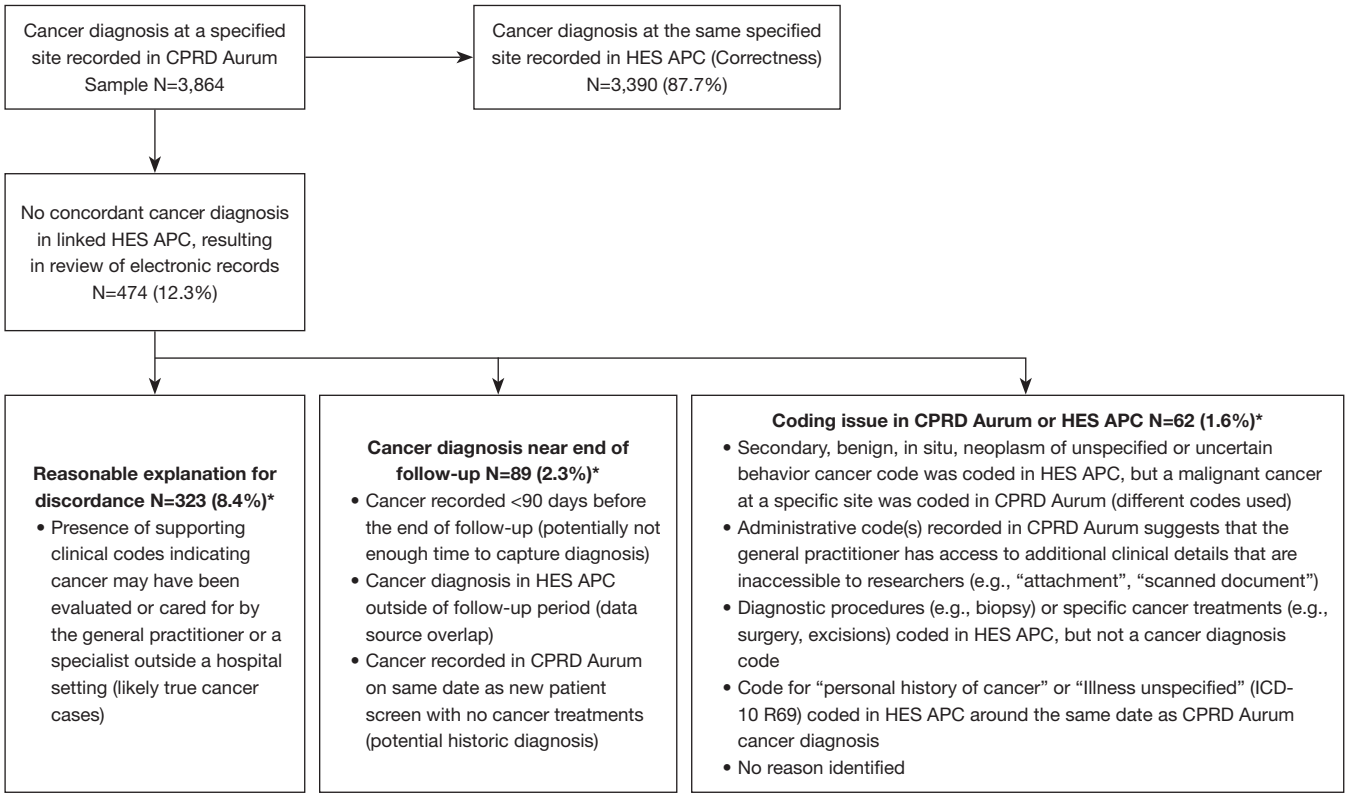

Figure 1 Reasons cancer diagnosis at a specific site may have been recorded in CPRD Aurum but not HES APC. CPRD, Clinical Practice Research Datalink; HES APC, Hospital Episode Statistics Admitted Patient Care. * Note: all proportions reported in figure among 3,864 cancer cases.

had supporting clinical codes.

Of the 3,864 cancer cases recorded in CPRD Aurum, we reviewed the electronic records for 474 (12.3%) with a diagnosis at a specified site recorded in CPRD Aurum and without a corresponding diagnosis code in HES APC to determine if there was a plausible reason for the discordant recordings (*Figure 1*). Among these 474 records reviewed, 323 had presence of cancer diagnosis plus supporting clinical codes recorded in CPRD Aurum, which may indicate that the cancer may have been under evaluation or cared for by the GP or a specialist outside a hospital setting (likely true cancer cases). Timing may have impacted the coding of cancer diagnoses and care received at the beginning or end of follow-up (89 of 474 records reviewed). There remained 62 of 474 records reviewed where coding issues in CPRD Aurum and/or HES APC may have explained the discordant recordings. Overall, 96.1% of the 3,864 patients with a cancer diagnosis at a specified site recorded in the CPRD Aurum sample had a concordant cancer diagnosis coded in HES APC (87.7%) or had a cancer diagnosis plus presence of supporting clinical codes recorded in CPRD Aurum indicating the cancer was cared for by a GP or specialist outside a hospital setting (8.4%).

Completeness of cancer diagnoses recorded in CPRD Aurum

There were 5,545 patients who had a code for cancer at a specified site recorded in HES APC, of which 3,390 (61.1%) also had a diagnosis recorded in CPRD Aurum at the same site (*Table 3*). Completeness estimates were lower early in the study period (1997–2004) and stabilized in later years (2005–2017) (*Table 3*). When stratified by age at first cancer diagnosis, completeness estimates were similar for those aged 20–49 (64.5%), 50–59 (69.0%), 60–69 (65.3%), and 70–79 (61.9%), but lower for those aged 80 years or older (50.2%) (*Table 3*). Completeness estimates were similar for females (61.0%) and males (61.2%) (*Table 3*).

Completeness estimates varied widely by cancer site (*Table 3*). Breast (86.5%) and male genital (84.0%) cancers

Table 3 Recording of cancer diagnoses in HES APC, compared to a sample of 50,000 patients in CPRD Aurum (completeness)

Category	Overall			Female			Male		
	N with code in HES APC sample	N with code in CPRD Aurum and HES APC	Completeness (%)	N with code in HES APC sample	N with code in CPRD Aurum and HES APC	Completeness (%)	N with code in HES APC sample	N with code in CPRD Aurum and HES APC	Completeness (%)
Overall[†]	5,545	3,390	61.1	2,635	1,608	61.0	2,910	1,782	61.2
By year of first cancer diagnosis									
1997–1999	500	205	41.0	245	101	41.2	255	104	40.8
2000–2004	1,144	678	59.3	557	332	59.6	587	346	58.9
2005–2009	1,373	898	65.4	643	400	62.2	730	498	68.2
2010–2014	1,542	994	64.5	737	481	65.3	805	513	63.7
2015–2017	986	615	62.4	453	294	64.9	533	321	60.2
By age at first cancer diagnosis (years)									
20–49	521	336	64.5	347	228	65.7	174	108	62.1
50–59	800	552	69.0	427	319	74.7	373	233	62.5
60–69	1,347	879	65.3	596	399	67.0	751	480	62.9
70–79	1,539	952	61.9	614	360	58.6	925	592	64.0
≥80	1,338	671	50.2	651	302	46.4	687	369	53.7
Malignant neoplasms, by site[‡]									
Lip, oral cavity and pharynx	108	61	56.5	37	18	48.7	71	43	60.6
Digestive organs	1,505	901	59.9	669	355	53.1	836	546	65.3
Colon	425	269	63.3	191	119	62.3	234	150	64.1
Liver	671	172	25.6	320	69	21.6	351	103	29.3
Pancreas	145	85	58.6	58	28	48.3	87	57	65.5
Rectum and Anus	308	185	60.1	118	74	62.7	190	111	58.4
Stomach, esophagus and small intestines	374	199	53.2	144	61	42.4	230	138	60.0
Respiratory and intrathoracic organs	989	453	45.8	427	189	44.3	562	264	47.0
Lung	904	422	46.7	393	181	46.1	511	241	47.2

Table 3 (*continued*)

Table 3 (*continued*)

Category	Overall			Female			Male		
	N with code in HES APC sample	N with code in CPRD Aurum and HES APC	Completeness (%)	N with code in HES APC sample	N with code in CPRD Aurum and HES APC	Completeness (%)	N with code in HES APC sample	N with code in CPRD Aurum and HES APC	Completeness (%)
Other respiratory and intrathoracic organs	161	19	11.8	62	NR	1.6	99	18	18.2
Bone and articular cartilage	467	32	6.9	178	15	8.4	289	17	5.9
Melanoma and other malignant neoplasms of skin	1,296	152	6.9	558	76	13.6	738	76	10.3
Mesothelial and soft tissue	287	31	10.8	162	11	6.8	125	20	16.0
Breast	719	622	86.5	712	618	86.8	7	NR	57.1
Female genital	307	182	59.3	307	182	59.3	n/a	n/a	n/a
Male genital organs	649	545	84.0	n/a	n/a	n/a	649	545	84.0
Prostate	615	518	84.2	n/a	n/a	n/a	615	518	84.2
Other	34	26	76.5	n/a	n/a	n/a	34	26	76.5
Urinary tract	472	241	51.1	126	58	46.0	346	183	52.9
Eye, brain and other parts of central nervous system	260	102	39.2	119	51	42.9	141	51	36.2
Thyroid and other endocrine glands	101	18	17.8	53	11	20.8	48	7	14.6
Lymphoid, hematopoietic and related tissues	410	259	63.2	189	120	63.5	221	139	62.9

[†], cancer was required to be at the same specified site. Metastatic, *in situ* and benign neoplasms, ill-defined and unspecified sites, and unspecified behavior were not evaluated in this study; [‡], patients may have cancer at more than one site. HES APC, Hospital Episode Statistics Admitted Patient Care; CPRD, Clinical Practice Research Datalink; NR, not reported due to cell counts of 1–4 people; n/a, not applicable.

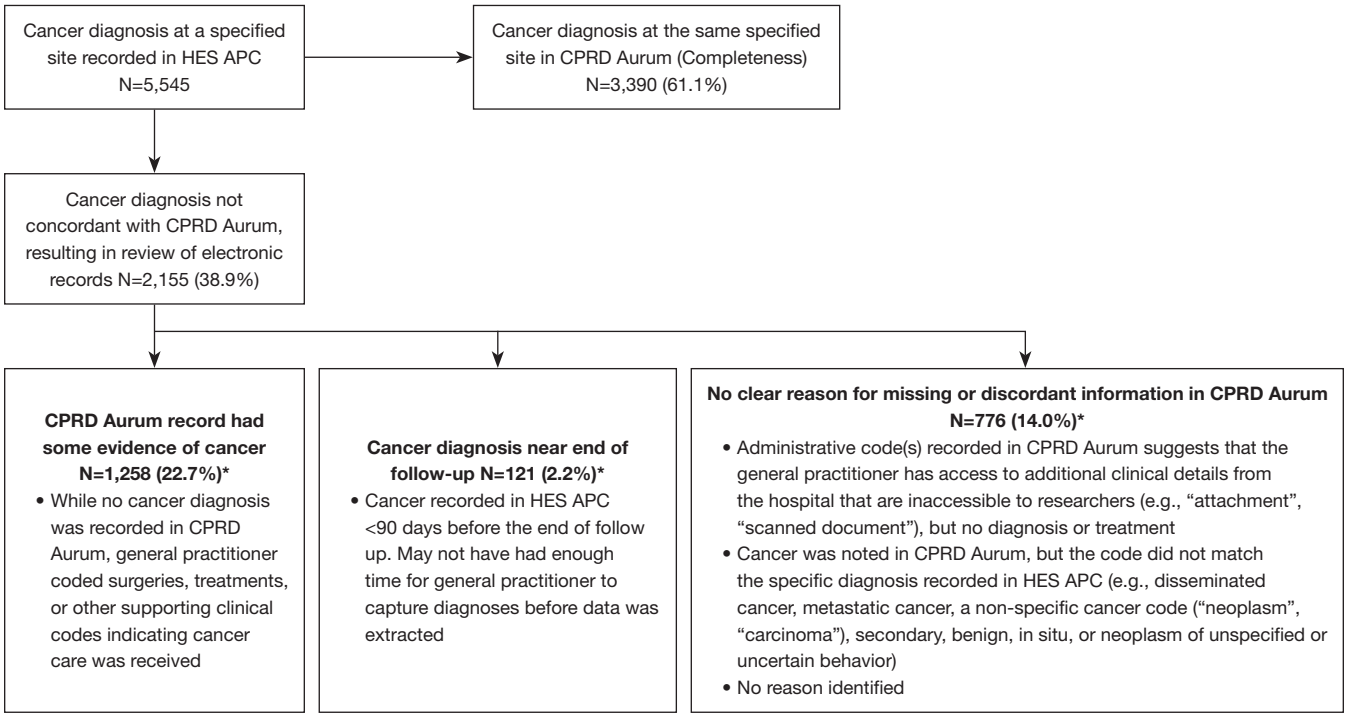

Figure 2 Reasons cancer diagnosis at a specified site were recorded in HES APC but not CPRD Aurum. HES APC, Hospital Episode Statistics Admitted Patient Care; CPRD, Clinical Practice Research Datalink. * Note: all proportions reported in figure among 3,864 cancer cases.

had the highest completeness. Completeness was lowest at sites typically associated with metastatic or secondary cancers: bone (6.9%), melanoma (6.9%), mesothelial and soft tissue (10.8%), other respiratory and intrathoracic organs (11.8%), and liver cancer (25.6%). Cancers of thyroid and other endocrine glands (17.8%) also had low completeness. Although completeness was similar by sex (61%) (*Table 3*), there were differences for some cancer sites: males had higher completeness for lip/oral cavity and digestive organs (60.6% *vs.* 48.7%), digestive organs (65.3% *vs.* 53.1%), mesothelial and soft tissues (16.0% *vs.* 6.8%), whereas females had higher completeness for thyroid and other endocrine gland cancers (20.8% *vs.* 14.6%).

From the 5,545 cancer cases in HES APC, we reviewed electronic records of 2,155 (38.9%) where there was a diagnosis code for cancer at a specified site recorded in the linked HES APC data without a corresponding diagnosis code recorded in CPRD Aurum to assess if there was a plausible reason for the discordant recordings (*Figure 2*). Among these 2,155 records reviewed, 1,258 had supporting clinical codes in CPRD Aurum indicating cancer care was received. The cancer diagnosis in HES APC was recorded near the end of the follow-up period in a further 121 patients of 2,155 reviewed, suggesting that there may not have been enough time for the GP to document the diagnosis or its care in CPRD Aurum. There remained 776 of 2,155 cancer cases in HES APC where there was no clear reason for missing or discordant information (*Figure 2*). Overall, 83.8% of 5,545 patients with a cancer diagnosis at a specified site coded in HES APC had a concordant diagnosis in CPRD Aurum (61.1%) or had presence of supporting clinical codes indicating cancer care was received (22.7%).

Discussion

The results of this study indicate that cancer diagnoses recorded in CPRD Aurum, where present, are of sufficient quality for most observational research. Throughout the study period (1997–2017), 87.7% of cancer diagnoses at a specified site recorded in CPRD Aurum were concordant with HES APC, while an additional 8.4% had a cancer diagnosis plus presence of supporting clinical codes recorded in CPRD Aurum indicating the cancer was cared

for by a GP or specialist (correctness). The completeness of cancer diagnosis recordings in CPRD Aurum compared with HES APC was 61.1%. An additional 22.7% of patients without the presence of a concordant diagnosis code in HES APC had other supporting clinical codes in CPRD Aurum where the GP indicated the patient had received cancer treatment and care. Completeness varied over the study period. Completeness estimates were higher for cancers at sites where GPs often prescribe ongoing drug therapy (e.g., breast cancer, prostate cancer), and were lower for cancers at sites typically associated with metastatic or secondary cancers (e.g., bone or articular cartilage, mesothelial and soft tissue), as well as for cancer sites that may be treated in outpatient specialist settings (e.g., melanoma, thyroid or other endocrine gland). Researchers should consider use of HES data linkage in addition to CPRD Aurum data if studying cancer sites where completeness estimates are low.

We chose to examine the coding of cancer diagnoses as part of our assessments of the CPRD Aurum data because cancer is a serious condition that requires medical attention and the patient care pathway spans both primary (GP) and secondary (hospital) healthcare settings. For these reasons, we expected that any patient who had a true cancer diagnosis would have a diagnosis recorded in both CPRD Aurum and HES APC data. However, it is likely that some patients with cancer received care in outpatient hospital settings or in specialist clinics, specifically for cancer sites that may be diagnosed or treated in a specialty care setting versus in hospital which could explain some of the low completeness numbers for certain cancer sites. It is likely that the increased concordance between the two data sources over calendar time, was due, at least in part, to more robust implementation of electronic data quality in the UK (16,17).

Where a cancer diagnosis was not recorded in both data sources, we looked for the presence of supporting clinical codes in CPRD Aurum indicating the patient received cancer treatment. The proportion of patients with a diagnosis of cancer at a specified site in HES APC that was not present in CPRD Aurum was as high as 38.9% when relying on codes for cancer diagnosis at specified sites to select cases; but missing cases were reduced to 15.5% when codes indicating cancer treatment and management were used to capture cases in CPRD Aurum. It is important to note that, given the presence of free text or administrative codes (e.g., "attachment", "scanned document", "letter"), GPs are likely aware of the patient's cancer status. GPs receive discharge letters detailing the various diagnoses and treatments received in hospital or specialist care settings, but GP staff must code these details into the electronic record for them to be available for use in research. Researchers should also consider using CPRD Aurum plus linked HES APC and Outpatient data, and/or linked cancer registry data from National Disease Registration Service to improve capture of cancer cases.

In this study, we required all CPRD Aurum patients selected for this random sample to have at least one admission for any reason in HES APC. This was necessary to have two data sources to compare. HES, in general, is not a perfect reference standard because coders may be non-clinical staff and there may be non-specific coding of some hospital events. In addition, some cancer events may be treated in outpatient hospital settings or non-NHS facility where some patients with private insurance may have opted for care elsewhere. We did not evaluate HES outpatient data in this study; therefore, correctness may be underestimated, particularly for cancers treated solely in outpatient hospital or other specialist cancer treatment settings. However, it is important to note that, unlike HES APC, it is not mandatory for diagnostic information to be recorded using ICD-10 codes in HES outpatient data and diagnostic information is captured in less than 5% of all attendances; therefore, the additional diagnostic information that could be provided by including HES outpatient data is likely to be small (18). It is also important to keep in mind that the goal of this study was to assess the quality of diagnosis recordings present in the CPRD Aurum data source, not to estimate unbiased measures of sensitivity and specificity. Cancer stage information is not available in CPRD Aurum or HES APC; therefore, we cannot assess differences in recording practices by cancer stage. Formal validation studies are still needed to assess the validity of cancer outcomes, including studies comparing CPRD Aurum to the Cancer Registry.

The results of this study are consistent with prior data quality assessments conducted in this same data sample and other UK primary care data (9-11,19,20). Diagnoses recorded in CPRD Aurum are of relatively high correctness for use in medical research, although completeness is variable across different cancers and over time and may not be sufficient for all research questions. Case capture could be improved by using linked data. Researchers should carefully consider study design, use of supporting clinical codes to enhance case selection, and use of linked data such as HES or cancer registry data to improve capture of cancer events.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the STROBE reporting checklist. Available at https://ace.amegroups.com/article/view/10.21037/ace-22-4/rc

Peer Review File: Available at https://ace.amegroups.com/article/view/10.21037/ace-22-4/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://ace.amegroups.com/article/view/10.21037/ace-22-4/coif). EY, TW and PM are employees of Clinical Practice Research Datalink (CPRD), the data custodians for CPRD Aurum. CPRD is jointly sponsored by the UK government's Medicines and Healthcare products Regulatory Agency and the National Institute for Health Research (NIHR). As a not-for-profit UK government body, CPRD seeks to recoup the cost of delivering its research services to academic, industry and government researchers through research user license fees. KWH, CV-S, RP and SJ are affiliated with Boston Collaborative Drug Surveillance Program (BCDSP). No funding was received by BCDSP for the conduct of this study. The BCDSP receives industry funding to conduct research using CPRD data.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was performed in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Independent Scientific Advisory Committee (ISAC) for Medicines and Healthcare Products Regulatory Agency (protocol No: 18_191). This study used anonymized electronic medical records, no patient contact occurred in its conduct.

References

1. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol 2019;48:1740-1740g.
2. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. BMJ 1991;302:766-8.
3. Jick SS, Kaye JA, Vasilakis-Scaramozza C, et al. Validity of the general practice research database. Pharmacotherapy 2003;23:686-9.
4. Herrett E, Thomas SL, Schoonen WM, et al. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol 2010;69:4-14.
5. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract 2010;60:e128-36.
6. Medicines & Healthcare Products Regulatory Agency. Clinical Practice Research Datalink [Internet]. Available online: www.CPRD.com [accessed 24 March 2021].
7. Walley T, Mantgani A. The UK General Practice Research Database. Lancet 1997;350:1097-9.
8. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol 2015;44:827-36.
9. Jick SS, Hagberg KW, Persson R, et al. Quality and completeness of diagnoses recorded in the new CPRD Aurum Database: evaluation of pulmonary embolism. Pharmacoepidemiol Drug Saf 2020;29:1134-40.
10. Persson R, Sponholtz T, Vasilakis-Scaramozza C, et al. Quality and Completeness of Myocardial Infarction Recording in Clinical Practice Research Datalink Aurum. Clin Epidemiol 2021;13:745-53.
11. Persson R, Vasilakis-Scaramozza C, Hagberg KW, et al. CPRD Aurum database: Assessment of data quality and completeness of three important comorbidities. Pharmacoepidemiol Drug Saf 2020;29:1456-64.
12. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20:144-51.
13. NHS Digital. Hospital Episode Statistics (HES) [Internet]. Available online: https://digital.nhs.uk/data-

and-information/data-tools-and-services/data-services/hospital-episode-statistics [accessed 24 March 2021].
14. Herbert A, Wijlaars L, Zylbersztejn A, et al. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int J Epidemiol 2017;46:1093-1093i.
15. International Statistical Classification of Diseases and Related Health Problems 10th Revision [Internet]. Available online: https://icd.who.int/browse10/2019/en [accessed 18 October 2019].
16. NHS Department of Health. A Simple Guide to Payment by Results. 2011. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213150/PbR-Simple-Guide-FINAL.pdf [accessed 24 March 2021].
17. Taggar JS, Coleman T, Lewis S, et al. 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. BMC Public Health 2012;12:329.
18. Hospital Episode Statistics (HES) Outpatient Care and CPRD primary care data documentation (set 21) Version 2.0. 2021. Available online: https://cprd.com/sites/default/files/2022-02/Documentation_HES_OP_set21.pdf [accessed 05 May 2022].
19. Strongman H, Williams R, Bhaskaran K. 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 2020;10:e037719.
20. Arhi CS, Bottle A, Burns EM, et al. Comparison of cancer diagnosis recording between the Clinical Practice Research Datalink, Cancer Registry and Hospital Episodes Statistics. Cancer Epidemiol 2018;57:148-57.

doi: 10.21037/ace-22-4
Cite this article as: Hagberg KW, Vasilakis-Scaramozza C, Persson R, Yelland E, Williams T, Myles P, Jick SS. Quality and completeness of malignant cancer recording in United Kingdom Clinical Practice Research Datalink Aurum compared to Hospital Episode Statistics. Ann Cancer Epidemiol 2022;6:6.

Supplementary

Appendix 1 Methods for estimating patient level start and end dates in CPRD Aurum calculated by Boston Collaborative Drug Surveillance Program (BCDSP)

Table S1 provides the definitions of variables used to estimate a patient-level StartDate and EndDate in CPRD Aurum electronic medical records. *Table S2* describes the steps of the StartDate algorithm and *Table S3* describes the steps of the EndDate algorithm.

Table S1 Definitions

Term	Definition
Plausible range	Date range when electronic medical records would be possibly in use. Estimated StartDate and EndDate must fall in this range. Minimum = January 1, 1988 (GPs first started using computers for electronic medical records in 1988) or January 1 of Year of Birth, whichever came later Maximum = date of CPRD Aurum data download
regstartdate	CPRD Aurum data field for most recent registration with practice (See CPRD Aurum Data Specification). This field may be missing or may contain implausible values.
regenddate	CPRD Aurum data field for the date the patient's registration ended with the practice (See CPRD Aurum Data Specification). This field may be missing or may contain implausible values.
Rx	Prescriptions (Issue table) in Plausible range
Clin	Clinical information that indicates that a patient is active in the record, such as diagnoses, vaccinations, labs, diagnostic, clinical care, referrals in Plausible range. Use MedCodes (Observation Table) with EMISCatID 1-5, 7, 9, 11-12, 14-16, 20-21, 25, 27-29, 32-36, and 39-47 in CPRD Aurum MedCode dictionary.
FirstRecordDate	First date of Rx or Clin in Plausible range
LastRecordDate	Last date of Rx or Clin in Plausible range
FirstRxDate	First date of Rx in Plausible range
FirstClinDate	Last date of Clin in Plausible range
DeathDate	There are two death date fields in CPRD Aurum (See CPRD Aurum Data Specification). At the time of this study, emis_ddate was used as the death date for estimation of a patient EndDate. Future studies will use cprd_ddate rather than emis_ddate.
lcd	CPRD Aurum data field for last data collection date (practice-level) (See CPRD Aurum Data Specification)

https://dx.doi.org/10.21037/ace-22-4

Table S2 BCDSP's algorithm to estimate a patient-level StartDate in CPRD Aurum

Algorithm step	Primary condition	Secondary condition	Action
1	regstartdate missing		StartDate = FirstRxDate
2	regstartdate missing	If no Rx	StartDate = FirstClinDate
3	regstartdate missing	If no Rx or Clin	StartDate = missing
4	regstartdate < Jan 1, 2000		StartDate = FirstRxDate
5	regstartdate < Jan 1, 2000	If no Rx	StartDate = FirstClinDate on or after regstartdate
6	regstartdate < Jan 1, 2000	If no Rx or Clin	Set StartDate to missing*
7	regstartdate ≥ Jan 1, 2000	If difference between FirstRxDate and regstartdate is ≤365 days	StartDate = regstartdate
8	regstartdate ≥ Jan 1, 2000	If difference between FirstRxDate and regstartdate is >365 days	StartDate = last of regstartdate or FirstRxDate
9	regstartdate ≥ Jan 1, 2000	If no Rx	StartDate = FirstClinDate on or after regstartdate
10	regstartdate ≥ Jan 1, 2000	If no Rx or Clin	Set StartDate to missing*
11	If StartDate in plausible range cannot be calculated by this point		Set StartDate to missing*

Note: Patterns of recording by GPs in electronic records changed in the early 2000's due to technological advances and the introduction of Quality and Outcome Framework (QOF), a pay-for-performance scheme which improved the collection of data due to reporting requirements. *Patients for whom a plausible StartDate cannot be estimated do not have clinical or prescription records of high quality for use in research.

Table S3 BCDSP's algorithm to estimate a patient-level EndDate in CPRD Aurum

Algorithm step	Primary condition	Secondary condition	Action
1	StartDate is missing*		Set EndDate to missing*
2	DeathDate is not missing		EndDate = minimum, not zero, of DeathDate and lcd
3	regenddate is not missing and within plausible range		EndDate = minimum, not zero, of regenddate and lcd
4	regenddate is missing	If difference between lcd and LastRecordDate is ≤365 days	EndDate = lcd
5	regenddate is missing	If difference between lcd and LastRecordDate is >365 days	EndDate = earliest of lcd and LastRecordDate
6	EndDate is calculated	if EndDate < StartDate	EndDate = StartDate
7	EndDate in plausible range cannot be calculated		Set EndDate to missing*

*Patients for whom a plausible StartDate and/or EndDate cannot be estimated do not have clinical or prescription records of high quality for use in research.

https://dx.doi.org/10.21037/ace-22-4