Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

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)

Katrina Wilcox Hagberg[1], Catherine Vasilakis-Scaramozza[2], Rebecca Persson[1], David Neasham[2], George Kafatos[2], Susan Jick[1,3]

[1]Epidemiology, Boston Collaborative Drug Surveillance Program, Lexington, MA, USA; [2]Center for Observational Research, Amgen Ltd, Uxbridge, UK; [3]Epidemiology, Boston University School of Public Health, Boston, MA, USA

Correspondence: Susan Jick, Boston Collaborative Drug Surveillance Program, 11 Muzzey Street, Lexington, MA, 02421, USA, Tel +1 781 862 6660, Fax +1 781 862 1680, Email sjick@bu.edu

Purpose: To evaluate the new Clinical Practice Research Datalink (CPRD) Aurum database, we estimated 'correctness' (ie accuracy, validity) and 'completeness' (ie presence, missingness) of malignant breast cancer diagnoses recorded in CPRD Aurum compared to external linked data sources: Hospital Episode Statistics (HES) Admitted Patient Care (APC), HES Outpatient (OP), and Cancer Registry (CR), and to the previously validated CPRD GOLD.

Methods: Linkage-eligible, female patients with incident malignant breast cancer diagnosis recorded in at least one study data source were selected. Correctness was the proportion of malignant breast cancer cases recorded in CPRD Aurum or GOLD who also had a diagnosis recorded in HES APC/OP (2004–2019) or CR (2004–2016). Completeness was estimated by identifying all malignant breast cancer diagnoses in HES APC/OP or CR and calculating the proportion with a concordant diagnosis in CPRD Aurum or GOLD.

Results: Compared to HES APC/OP, there were 85,659 and 31,452 eligible patients in CPRD Aurum and GOLD, respectively. Correctness estimates were high (CPRD Aurum 83.5%, GOLD 81.7%). Compared to CR, there were 70,190 and 29,597 eligible patients in CPRD Aurum and GOLD, respectively: correctness was 89.1% for CPRD Aurum and 88.2% for GOLD. Completeness estimates for CPRD Aurum and GOLD were high (>90%). Diagnoses were recorded in CPRD Aurum within −7 to 74 days of those in the linked sources. Reasons for discordant diagnostic coding included presence of treatment or other clinical codes only, diagnosis coded after end of follow-up, non-malignant breast cancer in linked data, and administrative codes in lieu of diagnostic codes.

Conclusion: These results indicate that correctness and completeness of malignant breast cancer diagnoses in CPRD Aurum were high and similar to CPRD GOLD. This provides confidence in use of CPRD Aurum for research purposes. Where complete case capture is important, researchers should consider linkage to HES APC or CR.

Keywords: CPRD Aurum, CPRD GOLD, breast cancer, validation

Introduction

The United Kingdom (UK) Clinical Practice Research Datalink (CPRD) has provided the well-described, widely used, and validated CPRD GOLD data for use in research for decades, though recently patient numbers have been in decline.[1] In 2018, CPRD made available a new primary care electronic database, CPRD Aurum. Since then, work to assess the quality and completeness of recording in CPRD Aurum has been undertaken to describe its characteristics, strengths, and limitations related to its suitability for use in medical research.[2–7] The results presented in our companion paper indicate that the recording of breast cancer diagnoses, treatments, and prescription drugs in CPRD Aurum were of similarly high quality compared to CPRD GOLD.[8] This study was conducted to extend the assessment to comparison with external linked data sources.

Received: 10 August 2023
Accepted: 29 November 2023
Published: 16 December 2023

The main objective of this study was to estimate ‘correctness’ (ie accuracy, validity) and ‘completeness’ (ie presence, missingness) of malignant breast cancer diagnoses recorded in CPRD Aurum, and separately CPRD GOLD, through comparison to external linked data sources: Hospital Episode Statistics (HES) Admitted Patient Care (APC), HES Outpatient (OP), and Cancer Registry data. We expect malignant breast cancer diagnoses to appear in the linked data sources because it often requires treatment in-hospital, reporting to the registry, and follow-up care by general practitioners (GPs) and consultants. The examination of recorded malignant breast cancer diagnoses in CPRD Aurum compared to those in external data linkages, as well as comparison of the same assessments in CPRD GOLD, will provide insight into the quality and completeness of information available in CPRD Aurum for research purposes. This study also describes in which of the data sources a malignant breast cancer diagnosis was recorded to help inform future research decisions.

Materials and Methods

Data Sources

CPRD Aurum and CPRD GOLD

CPRD Aurum and CPRD GOLD are two large, longitudinal, UK population-based electronic health record databases. The similarities and differences in electronic patient management systems, data structure, and geographical location between these two data sources have been previously described in detail.[1,8–10] In short, participating GPs record clinical data in an deidentified format, including demographic information, medical diagnoses, symptoms, referrals, details of hospital stays and specialist visits, and prescriptions. It is a requirement of the NHS that information regarding hospitalizations and outpatient referrals are sent to the GP, though GPs or their staff must code these details into the electronic record for them to be available for use in research. Henceforth, ‘GP record’ will be used to refer to the information coded in the patient’s CPRD Aurum or CPRD GOLD electronic medical record.

Hospital Episode Statistics (HES) Admitted Patient Care (APC) and Outpatient (OP)

We used two Hospital Episode Statistics (HES) linked datasets as external validation comparators for this study (1 January 2004–30 June 2019, the end of HES data at the time of download): 1) Admitted Patient Care (APC) and 2) Outpatient (OP) data.[11,12] The primary purpose of HES data is to enable hospitals to be paid for delivered care,[13] but the data are also used in multiple ways, including medical research and planning, monitoring trends and patterns in NHS hospital activity and providing the basis for national indicators of clinical quality and secondary care delivery. Nearly all CPRD Aurum and 56% for CPRD GOLD practices in England were linked to HES data at the time of data acquisition for this study.[13] HES APC data,[14] which captures inpatient hospitalizations in England since 1997, contain details of each hospital stay, including dates of admission and discharge, diagnoses (using ICD-10 codes),[15] and procedures performed (using OPCS codes).[16] Since 2003, HES OP data[17] has captured information on the type of outpatient consultation, appointment dates, the main specialty and specialty under which the patient was treated, and clinical diagnoses and procedures performed.[18] Unlike HES APC, it is not required that diagnostic information be recorded using ICD-10 codes in HES OP and diagnostic information is captured on less than 5% of all attendees.[17]

Cancer Registry Data from NHS Digital Disease Registration Service (NDRS)

Cancer data provided by NHS Digital National Disease Registration Service (NDRS)[19] via the National Cancer Registration and Analysis Service (NCRAS) Set 18 (henceforth referred to as Cancer Registry) was used as an external validation data source (1 January 2004–31 December 2016, the end of Cancer Registry data at the time of download).[20] These data are collected from multiple local cancer registries in England and, as such, there is variation in the data collected across contributors and over time. Information in the registry includes dates of cancer diagnoses (coded with ICD-10) and stage of cancer, though not all information is present for all people.[20] The Cancer Registry is a dynamic database, in which cancer registrations can be added or amended up to a year after the initial record.

Population Selection

For the HES comparison, we selected patients with a malignant breast cancer diagnosis who were eligible for HES linkage and present in the CPRD Aurum or CPRD GOLD data from 1 January 2004 through 30 June 2019. The patient's follow-up period was defined as the start of the patient's GP record or 1 January 2004, whichever came later, through the end of the patient's GP record or 30 June 2019, whichever came first. We excluded patients who had any of the following in the GP or HES records before the start of follow-up: malignant cancer diagnosis at any site (except non-melanoma skin cancer), mastectomy, or prescriptions for breast cancer drugs (eg tamoxifen, aromatase inhibitors).

For the Cancer Registry comparison, we selected patients with a malignant breast cancer diagnosis who were eligible for Cancer Registry linkage and present in the CPRD Aurum or CPRD GOLD data from 1 January 2004 through 31 December 2016. The patient's follow-up period was defined as the start of the patient's electronic GP record or 1 January 2004, whichever came later, through the end of the patient's electronic record or 31 December 2016, whichever came first. As with the HES comparison, patients with malignant cancer diagnosis at any site (except non-melanoma skin cancer), mastectomy, or prescriptions for breast cancer drugs (eg tamoxifen, aromatase inhibitors).

We selected patients in CPRD Aurum and CPRD GOLD with linkage to HES APC, HES OP, and Cancer Registry and who had a malignant breast cancer diagnosis recorded in at least one of the data sources during follow-up. Focus on this subgroup allowed us to describe which data sources contained malignant breast cancer diagnoses to help inform data source selection decisions in future studies. For this comparison, patient follow-up was defined as the start of the patient's GP record or 1 January 2004, whichever came later, through the end of the patient's GP or 31 December 2016, whichever came first (the period during which all data sources were present). We applied the same exclusions as for the HES and Cancer Registry comparisons.

Statistical Analysis

Using the data quality assessment methodologies described by Weiskopf and Weng[21] we estimated correctness (eg accuracy) and completeness (eg presence) of records in CPRD Aurum and CPRD GOLD, compared to HES APC/OP and the Cancer Registry. Correctness was calculated as the proportion of patients with a malignant breast cancer diagnosis in the GP record who also had a concordant diagnosis recorded in HES APC/OP, or the Cancer Registry, during follow-up. Completeness was calculated as the proportion of patients with a malignant breast cancer diagnosis in HES APC/OP, or the Cancer Registry, who also had a malignant breast cancer diagnosis recorded in CPRD Aurum or CPRD GOLD, separately, during follow-up. Among patients with a breast cancer diagnosis in the GP record and HES APC/OP or Cancer Registry, we also compared the recorded breast cancer diagnosis dates and reported median absolute difference in days (interquartile range (IQR)). Among patients who had a breast cancer diagnosis recorded in only one of the GP records or HES APC/OP or Cancer Registry (discordant records), we described reasons why the breast cancer diagnosis was recorded in only one data source (recorded ±365 days of breast cancer diagnosis date). Finally, among the subset of CPRD Aurum or CPRD GOLD patients with linkage to all data sources, we describe which data source(s) contained a malignant breast cancer diagnosis record. 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 are 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. Hospital Episode Statistics (HES) and NCRAS data Copyright © (2018), re-used with the permission of The Health & Social Care Information Centre. All rights reserved. This study was approved by Research Data Governance (RDG) (protocol no 20_000062), and the protocol was made available to the journal reviewers upon request.

Results

Comparison to HES APC/OP

There were 85,659 patients in CPRD Aurum and 31,452 patients in CPRD GOLD with a malignant breast cancer diagnosis who were eligible for the comparison with HES APC/OP (Supplementary Figure 1). Age at first breast cancer diagnosis and length of follow-up was similar for CPRD Aurum and CPRD GOLD. There were differences by region for

CPRD Aurum (higher proportion of GPs in North West and West Midlands) and CPRD GOLD (higher proportion of GPs in East of England and South East Coast). The year of first breast cancer diagnosis remained consistent over the study period for CPRD Aurum, whereas the proportion of breast cancer diagnoses in CPRD GOLD declined over the study period, reflecting the decline of GPs contributing to CPRD GOLD over time (Table 1). Median follow-up was 14.2 years (interquartile range (IQR) 8.4–15.5) for CPRD Aurum and 10.5 years (IQR 7.0–12.6) for CPRD GOLD.

Table 1 Characteristics of the Eligible Population and Presence of Breast Cancer Diagnoses, Treatments, and Other Supporting Clinical Codes, by Data Source

Characteristics	Comparison to HES APC/OP		Comparison to Cancer Registry	
	CPRD Aurum N=85,659 (%)	CPRD GOLD N=31,452 (%)	CPRD Aurum N=70,190 (%)	CPRD GOLD N=29,597 (%)
Age (years) at first breast cancer diagnosis				
<30	458 (0.5)	183 (0.6)	333 (0.5)	150 (0.5)
30–39	3677 (4.3)	1296 (4.1)	2982 (4.3)	1208 (4.1)
40–49	13,346 (15.6)	4979 (15.8)	11,045 (15.7)	4690 (15.9)
50–59	20,400 (23.8)	7402 (23.5)	16,428 (23.4)	6881 (23.3)
60–69	21,248 (24.8)	7957 (25.3)	17,521 (25.0)	7559 (25.5)
70–79	14,183 (16.6)	5072 (16.1)	11,330 (16.1)	4750 (16.1)
≥80	12,347 (14.4)	4563 (14.5)	10,551 (15.0)	4359 (14.7)
Mean ± st. dev. Median (IQR)	62.9 ± 14.5 62.4 (51.6–73.3)	62.9 ± 14.6 62.4 (51.5–73.4)	63.0 ± 14.6 62.5 (51.3–73.7)	63.0 ± 14.6 62.5 (51.6–73.5)
Year of first breast cancer diagnosis				
2004–2009	29,641 (34.6)	14,958 (47.6)	29,802 (42.5)	15,024 (50.8)
2010–2014	28,407 (33.2)	11,913 (37.9)	28,303 (40.3)	11,841 (40.0)
2015–2019 (2015-2016 for Cancer Registry)	27,611 (32.2)	4581 (14.6)	12,085 (17.2)	2732 (9.2)
Length of follow-up (years) Mean ± st. dev. Median (IQR)	11.8 ± 4.5 14.2 (8.4–15.5)	9.9 ± 4.0 10.5 (7.0–12.6)	10.2 ± 3.7 12.9 (7.8–13.0)	9.3 ± 3.5 10.3 (6.9–12.4)
Region of England				
North East	3891 (4.5)	578 (1.8)	3209 (4.6)	568 (1.9)
North West	14,585 (17.0)	4693 (14.9)	12,018 (17.1)	4458 (15.1)
Yorkshire and the Humber	3113 (3.6)	1002 (3.2)	2633 (3.8)	989 (3.3)
East Midlands	1978 (2.3)	706 (2.2)	1624 (2.3)	709 (2.4)
West Midlands	16,051 (18.7)	3495 (11.1)	13,210 (18.8)	3267 (11.0)
East of England	5394 (6.3)	3138 (10.0)	4428 (6.3)	3086 (10.4)
South West	12,040 (14.1)	3797 (12.1)	9827 (14.0)	3673 (12.4)
South Central	10,824 (12.6)	4179 (13.3)	8822 (12.6)	3986 (13.5)
London	10,340 (12.1)	4642 (14.8)	8284 (11.8)	4209 (14.2)
South East Coast	7443 (8.7)	5222 (16.6)	6135 (8.7)	4652 (15.7)

Abbreviations: IQR, Interquartile Range); st. dev., Standard Deviation.

Correctness compared to HES APC/OP (Table 2) was high and similar for CPRD Aurum (83.5%) and CPRD GOLD (81.7%) and remained >80% when stratified by time period. Correctness estimates in CPRD Aurum and CPRD GOLD were greater than 80% for age groups between 30–79 years. Correctness was 84.0% and 71.9% for patients <30 and ≥80 years in CPRD Aurum, respectively, whereas for CPRD GOLD, correctness was 69.6% and 67.0% for patients <30 and ≥80 years, respectively (Table 2).

Completeness was high and similar for CPRD Aurum (91.5%) and CPRD GOLD (90.0%) and remained high when stratified by time period (Table 2). Completeness estimates were lowest for the youngest and highest age groups in both CPRD Aurum and CPRD GOLD (Table 2).

Among patients who had a malignant breast cancer diagnosis recorded in both data sources, we calculated the number of days between the first malignant breast cancer diagnosis date recorded in HES APC or HES OP compared to the first present in the GP record. The median differences between HES APC and CPRD Aurum was 21 days (IQR 10–35) and 20 days (IQR 8–32) between HES APC and CPRD GOLD. Compared to HES OP, the median differences were 74 days (IQR 28–194) for CPRD Aurum and 67 days (IQR 28–183) in CPRD GOLD.

Comparison to the Cancer Registry

There were 70,190 patients in the CPRD Aurum cohort and 29,597 patients in CPRD GOLD with malignant breast cancer who were eligible for the comparison with the Cancer Registry (Supplementary Figure 2). Age at first breast cancer diagnosis and length of follow-up were similar for CPRD Aurum and CPRD GOLD. Similar to the HES comparison, there were differences between the two cohorts in terms of region and year of first breast cancer diagnosis (Table 1).

Correctness compared to the Cancer Registry (Table 3) was high and similar for CPRD Aurum (89.1%) and CPRD GOLD (88.2%) and remained high (>87%) when stratified by calendar period. Correctness estimates in CPRD Aurum were high (>80%) for all age groups. For CPRD GOLD, correctness was high (85% or higher) for ages 30 years and older, and 78.1% for patients aged <30 years.

Completeness was high and similar for CPRD Aurum (92.4%) and CPRD GOLD (91.2%) and remained greater than 90% when stratified by calendar period, except in the most recent calendar period in CPRD GOLD (2015–2016) where completeness was 86.3% (Table 3). Completeness in CPRD Aurum was high (>85%) for all age groups, while for CPRD GOLD completeness was high (85% or higher) for those aged ≥30 years, and 74.5% for patients aged <30 years. Breast cancer stage information is only available in the Cancer Registry. Completeness of breast cancer case information in CPRD Aurum and CPRD GOLD was greater than 80% for all cancers that could be staged (Stage 0–4) and lower for cancers that were unstageable (Table 3).

Among patients who had a malignant breast cancer diagnosis recorded in both the GP record and the Cancer Registry, we calculated the number of days between the first malignant breast cancer diagnosis date recorded in the Cancer Registry compared to the first present in the GP record. The median difference in the recorded diagnosis dates was −7 days (IQR −14 to −1) for CPRD Aurum and −8 days (IQR −14 to −2) for CPRD GOLD, indicating that the date in the GP record occurred first.

Reasons for Breast Cancer Diagnosis in Only One Data Source

Among patients who had a breast cancer diagnosis recorded in only one of the data sources (discordant records), the reasons for this were similar for CPRD Aurum and CPRD GOLD (Table 4). Common reasons included having codes for treatment of breast cancer (eg mastectomy, lumpectomy, drug prescription, radiation, chemotherapy) or other supporting clinical codes (eg lump, suspected breast cancer, referrals, cancer care or palliative care) in the GP record in the absence of a malignant breast cancer diagnosis code or having a diagnosis coded after the end of patient follow-up (timing of coding). Another common reason was discordant coding, where patients had a malignant breast cancer code in one data source and a non-malignant breast cancer (eg in situ, benign) coded in the linked data. Some patients had administrative codes indicating receipt of a ‘letter’ or ‘scanned document’ from hospitals or specialists that may have contained diagnostic information that was not subsequently coded in the patient’s GP record. In the HES comparison a similar proportion of patients in CPRD Aurum and CPRD GOLD had malignant breast cancer coded in the GP record but had non-specific codes in the HES OP record. In the Cancer Registry comparison, a similar proportion of patients in CPRD

Table 2 Correctness and Completeness of Malignant Breast Cancer Diagnosis Recorded in CPRD Aurum and CPRD GOLD Compared to HES APC/HES OP

Correctness	**CPRD Aurum**			**CPRD GOLD**		
	Breast Cancer in CPRD Aurum (N)	**Breast Cancer in CPRD Aurum and HES APC/ OP (N)**	**CPRD Aurum Correctness Estimate**	**Breast Cancer in CPRD GOLD (N)**	**Breast Cancer in CPRD GOLD and HES APC/OP (N)**	**CPRD GOLD Correctness Estimate**
Correctness Overall	79,481	66,394	83.5%	28,825	23,549	81.7%
Correctness by year of breast cancer diagnosis						
2004–2009	27,570	23,115	83.8%	13,848	11,276	81.4%
2010–2014	26,389	22,513	85.3%	10,846	8940	82.4%
2015–2019	25,522	20,766	81.4%	4131	3333	80.7%
Correctness by age at breast cancer diagnosis						
<30	399	335	84.0%	148	103	69.6%
30–39	3413	2966	86.8%	1209	1023	84.6%
40–49	12,562	10,761	85.7%	4649	3910	84.1%
50–59	19,196	16,036	83.5%	6874	5624	81.8%
60–69	19,817	17,014	85.9%	7328	6232	85.0%
70–79	13,042	11,336	86.8%	4610	3973	86.2%
≥80	11,052	7946	71.9%	4007	2684	67.0%

Completeness	Breast Cancer in HES APC/OP (N)	Breast Cancer in CPRD Aurum and HES APC/OP (N)	CPRD Aurum Completeness Estimate	Breast Cancer in CPRD GOLD (N)	Breast Cancer in CPRD GOLD and HES APC/OP (N)	CPRD GOLD Correctness Estimate
Completeness Overall	75,572	66,394	91.5%	26,176	23,549	90.0%
Completeness by year of breast cancer diagnosis						
2004–2009	24,111	22,162	91.9%	11,933	10,869	91.1%
2010–2014	24,557	22,548	91.8%	10,261	9179	89.5%
2015–2019	23,904	21,684	90.7%	3982	3501	87.9%
Completeness by age at breast cancer diagnosis						
<30	380	321	84.5%	136	103	75.7%
30–39	3146	2895	92.0%	1073	990	92.3%
40–49	11,364	10,606	93.3%	4189	3862	92.2%
50–59	17,253	16,046	93.0%	6167	5642	91.5%
60–69	18,426	17,015	92.3%	6854	6224	90.8%
70–79	12,490	11,318	90.6%	4415	3948	89.4%
≥80	9513	8193	86.1%	3342	2780	82.3%

Abbreviations: APC, Admitted Patient Care; CPRD, Clinical Practice Research Datalink; HES, Hospital Episode Statistics; N, number of patients; OP, Outpatient.

Table 3 Correctness and Completeness of Malignant Breast Cancer Diagnosis Recorded in CPRD Aurum and CPRD GOLD Compared to the Cancer Registry

Correctness	CPRD Aurum			CPRD GOLD		
	Breast Cancer in CPRD Aurum (N)	**Breast Cancer in CPRD Aurum and Cancer Registry (N)**	**CPRD Aurum Correctness Estimate**	**Breast Cancer in CPRD GOLD (N)**	**Breast Cancer in CPRD GOLD and Cancer Registry (N)**	**CPRD GOLD Correctness Estimate**
Correctness Overall	65,377	58,233	89.1%	27,276	24,049	88.2%
Correctness by year of breast cancer diagnosis						
2004–2009	27,691	24,520	88.6%	13,937	12,193	87.5%
2010–2014	26,406	23,699	89.8%	10,868	9680	89.1%
2015–2016	11,280	10,014	88.8%	2471	2176	88.1%
Correctness by age at breast cancer diagnosis						
<30	302	261	86.4%	137	107	78.1%
30–39	2805	2563	91.4%	1124	991	88.2%
40–49	10,416	9370	90.0%	4387	3910	89.1%
50–59	15,641	13,442	85.9%	6448	5482	85.0%
60–69	16,514	14,530	88.0%	7004	6122	87.4%
70–79	10,496	9631	91.8%	4343	3983	91.7%
≥80	9203	8436	91.7%	3833	3454	90.1%

Completeness	Breast Cancer in Cancer Registry (N)	Breast Cancer in CPRD Aurum and Cancer Registry (N)	CPRD Aurum Completeness Estimate	Breast Cancer in CPRD GOLD (N)	Breast Cancer in CPRD GOLD and Cancer Registry (N)	CPRD GOLD Completeness Estimate
Completeness Overall	63,046	58,233	92.4%	26,370	24,049	91.2%
Completeness by year of breast cancer diagnosis						
2004–2009	26,362	24,453	92.7%	13,196	12,210	92.5%
2010–2014	25,737	23,872	92.8%	10,710	9712	90.7%
2015–2016	10,947	9908	90.5%	2464	2127	86.3%
Completeness by age at breast cancer diagnosis						
<30	287	261	90.8%	145	108	74.5%
30–39	2719	2562	94.5%	1066	995	93.3%
40–49	9970	9421	94.5%	4203	3927	93.4%
50–59	14,215	13,393	94.2%	5902	5466	92.6%
60–69	15,538	14,542	93.6%	6680	6138	91.9%
70–79	10,490	9615	91.7%	4386	3966	90.4%
≥80	9827	8439	85.9%	3988	3449	86.5%
Completeness by Breast Cancer Stage Best						
Stage 0	299	255	85.3%	46	40	87.0%
Stage 1	17,773	16,702	94.0%	3196	2935	91.8%
Stage 2	14,870	13,947	93.8%	2789	2564	91.9%
Stage 3	3186	3001	94.2%	575	530	92.2%
Stage 4	2271	1875	82.6%	413	339	82.1%
Unstageable	93	64	68.8%	17	10	58.8%
Insufficient information	2053	1715	83.5%	278	231	83.1%
Missing stage	22,501	20,674	91.9%	5204	4775	91.8%

Abbreviations: CPRD, Clinical Practice Research Datalink; N, number of patients.

Table 4 Reasons Why Malignant Breast Cancer Cases Were Found in Only One Data Source (CPRD Aurum or HES APC/HES OP)

Reasons[a]	Comparison to HES APC / HES OP				Comparison to Cancer Registry			
	CPRD Aurum		CPRD GOLD		CPRD Aurum		CPRD GOLD	
	In CPRD Aurum Only N=13,087 (%)	In HES APC/ HES OP only N=6178 (%)	In CPRD GOLD Only N=5276 (%)	In HES APC/ HES OP only N=2627 (%)	In CPRD Aurum Only N=7144 (%)	In Cancer Registry only N=4813 (%)	In CPRD GOLD Only N=3227 (%)	In Cancer Registry only N=2321 (%)
Non-malignant breast cancer code recorded (eg in situ, benign, history of BrCa) (discordant coding)	4816 (36.8)	2944 (41.2)	1721 (32.6)	531 (20.2)	4,622 (64.7)	1840 (38.2)	1906 (59.1)	496 (21.4)
Codes for other cancer site(s) present	534 (4.1)	948 (15.3)	222 (4.2)	524 (20.0)	n/a	627 (13.0)	n/a	433 (18.7)
Treatment for breast cancer (eg mastectomy, lumpectomy, drug prescription, radiation, chemotherapy) coded in CPRD Aurum / CPRD GOLD	n/a	3856 (62.4)	n/a	1563 (59.5)	n/a	3077 (63.9)	n/a	1535 (66.1)
Sign of breast cancer (eg lump, suspected, biopsy) or supporting clinical code (eg referral, cancer care, palliative care) coded in CPRD Aurum / CPRD GOLD	n/a	4281 (69.3)	n/a	1799 (68.5)	n/a	3512 (73.0)	n/a	1781 (76.7)
Malignant breast cancer diagnosis recorded in other data source just after end of follow-up (timing of coding)	1340 (10.2)	1,239 (20.1)	437 (8.3)	607 (23.1)	n/a	1366 (28.4)	n/a	576 (24.8)
Uninformative administrative code (eg "scanned document", "letter") present in CPRD Aurum / CPRD GOLD	n/a	5788 (94.7)	n/a	2091 (79.6)	n/a	4393 (91.3)	n/a	1772 (76.4)
HES OP ICD-10 code "R69.X6 Unknown and Unspecified Causes of Morbidity" and/or OPCS code "X99.7 Procedure carried out, but no appropriate OPCS-4 code" present, but not other diagnosis or procedure codes present in HES OP	11,813 (90.3)	n/a	4638 (87.9)	n/a	n/a	n/a	n/a	n/a
Patient eligible for CR linkage, but did not have any breast cancer diagnosis (malignant or in situ) in Cander Registry	n/a	n/a	n/a	n/a	2481 (34.7)	n/a	1279 (39.6)	n/a
No reason identified	989 (7.6)	272 (4.4)	547 (10.4)	6 (0.2)	30 (0.4)	253 (5.3)	36 (1.1)	115 (5.0)

Notes: [a]Recorded ±365 days of breast cancer diagnosis date. Patients can have more than one reason.

Abbreviations: APC, Admitted Patient Care; CPRD, Clinical Practice Research Datalink; HES, Hospital Episode Statistics; N, number of patients; OP, Outpatient.

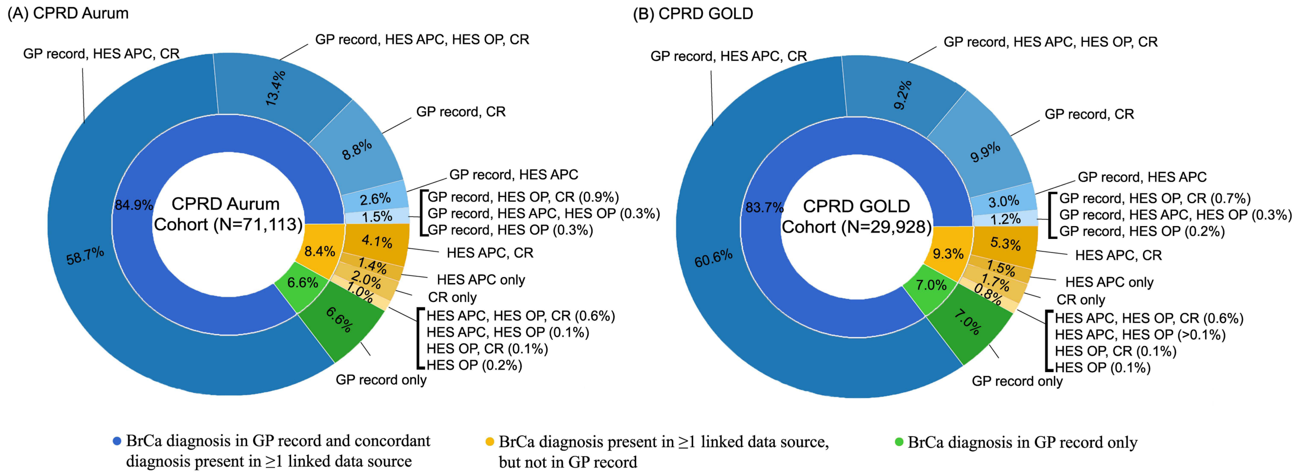

Figure 1 Comparison of malignant breast cancer diagnosis recorded in the CPRD Aurum (**A**) or CPRD GOLD (**B**) GP record, HES APC, HES OP, and Cancer Registry among patients who are eligible for linkage and who had at least 1 malignant breast cancer diagnosis recorded during follow-up.
Abbreviations: APC, Admitted Patient Care, CPRD, Clinical Practice Research Datalink; CR, Cancer Registry, GP, general practitioner; HES, Hospital Episode Statistics; OP, Outpatient.

Aurum and CPRD GOLD had a malignant breast cancer coded in the GP record but did not have any breast cancer diagnosis (malignant or in situ) in the Cancer Registry.

Presence of Breast Cancer Diagnosis by Data Source

We investigated in which data source(s) patients were likely to have a malignant breast cancer diagnosis recorded to help inform decisions regarding data source selection (Figure 1). We selected a subgroup of patients in CPRD Aurum (N=71,113, Figure 1A) and CPRD GOLD (N=29,928, Figure 1B) who were eligible for linkage to HES APC, HES OP, and the Cancer Registry who had a malignant breast cancer diagnosis in one or more of the data sources during follow up and were free from evidence of prior breast cancer. Overall, most patients had a malignant breast cancer diagnosis recorded in the GP record (91.6% in CPRD Aurum, 90.7% in CPRD GOLD). Approximately 81.1% of patients in CPRD Aurum also had a diagnosis in HES APC compared to 80.4% in CPRD GOLD. A lower proportion of patients had a diagnosis in HES OP (15.8% in CPRD Aurum, 11.1% in CPRD GOLD). Approximately 88.6% and 88.0% of patients in CPRD Aurum and CPRD GOLD, respectively, had a malignant breast cancer diagnosis recorded in the Cancer Registry.

When we compared the presence of a malignant breast cancer diagnosis in each data source (GP record, HES AP/OP, and Cancer Registry) (Figure 1), most patients had a diagnosis recorded in the GP record and in HES APC/HES OP and/or Cancer Registry (84.9% of the CPRD Aurum cohort and 83.7% in the CPRD GOLD cohort). Most of these (58.7% of the CPRD Aurum, 60.6% of the CPRD GOLD cohorts) had a diagnosis in the GP record, HES APC, and the Cancer Registry (not in HES OP). A breast cancer diagnosis was coded in one of the linked data sources (HES APC/OP or Cancer Registry) but did not have a diagnosis in the GP record for 8.4% of patients in CPRD Aurum and 9.3% in CPRD GOLD cohort. Approximately 6.6% of patients in CPRD Aurum and 7.0% of patients in CPRD GOLD had a malignant breast cancer diagnosis recorded in the GP record only. While it is possible that patients missing concordant diagnoses in HES or CR represent provisional breast cancer recorded in the GP record, most (92.9% Aurum, 89.0% GOLD) had codes for relevant treatments or care present in their GP record that supported a breast cancer diagnosis. Median age of breast cancer cases recorded in linked data only was higher (HES APC/OP 67 years, Cancer Registry 68 years) than cases recorded only in GP data (60 years).

Discussion

The results of this study indicate that, compared to external linked hospital and cancer registry data, malignant breast cancer diagnoses recorded in CPRD Aurum, where present, are of sufficient quality for most observational research and similar to the well-described CPRD GOLD. The completeness of malignant breast cancer diagnosis recordings in CPRD Aurum compared with HES APC/OP and the Cancer Registry were also high and similar to CPRD GOLD.

Our comparison of the presence of malignant breast cancer diagnoses recorded in all linked data sources indicate that researchers may increase capture of malignant breast cancer cases by using linked HES or Cancer Registry data (by 8.4% in CPRD Aurum, 9.3% in CPRD GOLD). Most of the cases missing from CPRD Aurum or CPRD GOLD data could be found using HES APC data, with a small proportion found in only the Cancer Registry. HES OP data provided few additional cases in the absence of the other linked data sources. It has been previously reported that using CPRD GOLD alone may result in selection of younger, healthier patients,[22,23] while use of HES APC alone over-represents older patients and those diagnosed through emergency routes,[23] a similar finding to this study. Given the similarities between the two primary care databases[8] these may also be a concern for CPRD Aurum data.

This study aimed to validate CPRD Aurum data using an oncology indication. We chose to evaluate the coding of malignant breast cancer diagnoses for several reasons: 1) breast cancer is a serious condition that requires medical attention, 2) the UK has a national breast cancer detection and prevention program and NICE guidelines that guide diagnosis and care, and 3) 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 malignant breast cancer diagnosis would have a diagnosis recorded in their CPRD Aurum or GOLD record, as well as in the linked HES and Cancer Registry external data comparators. This provided an opportunity to estimate the correctness and completeness of the coding of malignant breast cancer.

We chose to use the well-established CPRD GOLD, an electronic general practice database similar to CPRD Aurum, as a comparator in this study given its history of prior validation evaluations. Our correctness and completeness estimates for CPRD GOLD are well aligned to those reported in prior publications.[22–25] Therefore, the CPRD GOLD results in our study provide a strong base against which to evaluate the quality of malignant breast cancer diagnoses recorded in CPRD Aurum. There have been two prior published studies looking at the presence of any cancer diagnoses (including breast cancer) in CPRD Aurum, though our current study expands on the number of data sources used for comparison (adding HES OP, Cancer Registry, and CPRD GOLD),[6] the amount of detail about breast cancer diagnosis and treatments, and our population is more generalizable as it is not limited to a selected population matched to patients with psoriasis.[22] That said, our estimates for correctness and completeness of malignant breast cancer recordings in CPRD Aurum were similar to those two prior evaluations of cancer, providing further confidence in the quality of the data. Finally, the timing of breast cancer diagnosis dates in our study was also similar to prior studies for both CPRD GOLD[22–24] and CPRD Aurum,[24] indicating breast cancer diagnoses recorded in the CPRD records are largely recorded within 3 months of those in HES or the Cancer Registry. Any small differences between prior publications and our results may be due to differences in use of different linked data sources, calendar periods, diagnostic code lists, screening programs, and population selection criteria. Overall, our results, combined with previously published studies, provide reassurance of the quality and completeness of breast cancer diagnoses in CPRD Aurum and that breast cancer diagnoses recorded in CPRD Aurum is of similar quality and completeness to those recorded in the previously validated CPRD GOLD.

Considerations for the Use of CPRD Aurum, CPRD GOLD, and Linkages

CPRD Aurum and CPRD GOLD data, both UK population-based electronic health databases, contain rich information including medical diagnoses, symptoms, referrals, demographic information, lab data, and prescriptions. The available linkages further expand on the types of data available. Researchers can choose to link to HES data (HES APC and OP) to gain access to diagnoses and procedures conducted in hospital (but not medications dispensed in hospital), while linkage to the Cancer Registry provides information on dates of cancer diagnoses and stage of cancer. Investigators can choose which CPRD and linked data source(s) to use based on their specific research question.

While CPRD electronic health data is a rich source of health information, researchers using CPRD Aurum and CPRD GOLD should be aware of changes to collection practices, policies, and software, which may result in changes to the quality and completeness diagnoses over time.[22] It is also important to recognize that the linked data sources used in this study are not perfect reference standards, nor do they contain all the types of data required for all research questions. For example, HES data may be coded by non-clinical staff and non-specific coding of hospital events is sometimes used. In addition, some cancer events may be treated in non-NHS facilities using private insurance and not captured in HES data. Unlike HES APC, it is not mandatory for diagnostic information to be recorded using ICD-10 codes in HES OP data and diagnostic information is captured in less than 5% of all attendances; therefore, as our evaluation shows, few additional breast cancer cases were identified by linking to HES OP data.[17]

The Cancer Registry, a dynamic data source which contains a record for each registerable tumor diagnosed and treated in England and reported to NDRS (formerly NCRAS), is often assumed by researchers to be a "gold standard". While it is an important resource, the completeness of data fields in the Cancer Registry varies significantly by tumor type and calendar time.[19] The Cancer Registry began as eight regional cancer registries, which merged into a single national registry in 2013 with initiatives to improve and standardize the capture of cancer data, including stage and grade.[23,26,27] Strongman et al did not use the Cancer Registry as an outright gold standard in their cancer algorithm because it is plausible that the data does not capture all tumors diagnosed and treated in primary or secondary care settings.[23] In our study, among those with linkages to all data sources, there were some malignant breast cancer cases recorded in HES or in the CPRD record that were not present in the Cancer Registry (11.5% CPRD Aurum, 12.1% CPRD GOLD). It is possible these missing cases in the Cancer Registry could be provisional diagnoses coded in the GP or HES record, though most of these patients had codes for relevant treatments or care in their GP record that supported the diagnosis. These patients may have non-malignant breast cancer coded, or these could also represent true malignant breast cancer cases that are missing from the Cancer Registry. The Cancer Registry is the only data source that contains information on stage or grade; therefore, studies that require data on cancer staging would need to link to the Cancer Registry. Completeness of stage information in the Cancer Registry has been reported to be lower for older age groups, patients diagnosed through an emergency route, those who died within 30 days of diagnosis, those who did not have a record of cancer related surgery, and those who had concomitant conditions.[26]

Researchers using CPRD data should also weigh the benefits of using linked data sources to address their specific study question against the impact of reduced study sample size (limited to patients eligible for linkage), length of patient follow-up, lags in availability of linked data (at time of publication, Cancer Registry data was available through 2018, HES through 2021), and geographic generalizability (restricted to England only). In addition, there are some significant administrative considerations, including additional costs of linkages and impact on study timeline given additional approvals and greater administrative complexities to receive a data cut (CPRD website indicates the process to access Cancer Registry data currently takes 12–18 months to complete).

Conclusion

Overall, the results of this study indicate that, compared to external linked hospital and cancer registry data sources, the correctness and completeness of breast cancer diagnoses recorded in CPRD Aurum were high and similar to CPRD GOLD. This provides confidence in the use of CPRD Aurum for research on malignant breast cancer. However, where complete case capture is important for specific research questions, researchers should consider linkage to HES APC or Cancer Registry.

Funding

This study was funded by Amgen, Ltd.

Disclosure

Ms Katrina Wilcox Hagberg, Dr Catherine Vasilakis-Scaramozza, Ms Rebecca Persson, and Susan Jick report the Boston Collaborative Drug Surveillance Program (employer) received contractual research grants from Amgen during the conduct of the study. Dr David Neasham and Dr George Kafatos are employees of Amgen Ltd and own shares of Amgen Inc. Dr Susan Jick reports she is a member of the Amgen Methods Council. The authors declare no other conflicts of interest in this work.

References

1. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44 (3):827–836. doi:10.1093/ije/dyv098
2. Jick S, 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(9):1134–1140. doi:10.1002/pds.4996

3. 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(11):1456–1464. doi:10.1002/pds.5135
4. Persson R, Sponholtz T, Vasilakis-Scaramozza C, et al. Quality and Completeness of Myocardial Infarction Recording in Clinical Practice Research Datalink Aurum. *Clinical Epidemiology*. 2021;13:745–775. doi:10.2147/CLEP.S319245
5. Persson R, Hagberg KW, Vasilakis-Scaramozza C, et al. Presence of codes for indication for use in Clinical Practice Research Datalink Aurum: an assessment of benign prostatic hyperplasia treatments. *Clin Epi*. 2022;641–652. doi:10.2147/CLEP.S360843
6. Hagberg KW, Vasilakis-Scaramozza C, Persson R, et al. Quality and completeness of malignant cancer recording in United Kingdom Clinical Practice Research Datalink Aurum compared to Hospital Episode Statistics. *Annals of Cancer Epi*. 2022;6:6. doi:10.21037/ace-22-4
7. Vasilakis-Scaramozza C, Hagberg KW, Persson R, et al. Quality of rheumatoid arthritis recording in United Kingdom Clinical Practice Research Datalink Aurum. *Pharmacoepidemiol Drug Saf*. 2023;32(1):73–77. doi:10.1002/pds.5551
8. Hagberg KW, Vasilakis-Scaramozza C, Persson R, Neasham D, Kafatos G, Jick S. Presence of Breast Cancer Information Recorded in United Kingdom Primary Care Databases: Comparison of CPRD Aurum and CPRD GOLD (Companion Paper 1). *Clin Epidemiol*. 2023;15:1183–1192. doi:10.2147/CLEP.S434795
9. Lawson DH, Sherman V, Hollowell J. The General Practice Research Database. Scientific and Ethical Advisory Group. *QJM*. 1998;91(6):445–452. doi:10.1093/qjmed/91.6.445
10. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epi*. 2019;48(6):1740–1740g. doi:10.1093/ije/dyz034
11. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol*. 2017;46(4):1093–1093i. doi:10.1093/ije/dyx015
12. Thorn JC, Turner E, Hounsome L, et al. Validation of the Hospital Episode Statistics Outpatient Dataset in England. *Pharmacoeconomics*. 2016;34 (2):161–168. doi:10.1007/s40273-015-0326-3
13. NHS Digital. Hospital Episode Statistics (HES). Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics. Accessed March 24, 2021.
14. CPRD. Hospital Episode Statistics (HES) Admitted Patient Care (APC) data dictionary. Available from: https://cprd.com/sites/default/files/2022-02/Data_Dictionary_HES_APC.pdf. Accessed March 24, 2021.
15. NHS Digital. NHS Classifications ICD-10. Available from: https://isd.digital.nhs.uk/trud/users/guest/filters/0/categories/28. Accessed March 24, 2021.
16. NHS Digital. OPCS Classification of Interventions and Procedures. Available from: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dapb0084-opcs-classification-of-interventions-and-procedures. Accessed 24, March 2021.
17. CPRD. Hospital Episode Statistics (HES) Outpatient Care and CPRD primary care documentation. Available from: https://cprd.com/sites/default/files/2022-02/Documentation_HES_OP_set21.pdf. Accessed March 24, 2021.
18. CPRD. CPRD linked data. Available from: https://cprd.com/cprd-linked-data. Accessed March 24, 2021.
19. NHS Digital. National Disease Registration Service. Available from: https://digital.nhs.uk/services/national-disease-registration-service. Accessed December 1, 2021.
20. Henson KE, Elliss-Brookes L, Coupland VH, et al. Data resource profile: National Cancer Registration Dataset in England. *Int J Epi*. 2020;49:16–16h. doi:10.1093/ije/dyz076
21. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Informatics Assoc*. 2013;20(1):144–151. doi:10.1136/amiajnl-2011-000681
22. Trafford AM, Parisi R, Rutter MK, Kontopantelis E, Griffiths CEM. Ashcroft D on behalf of the Global Psoriasis Atlas. 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*. 2021;16(7):e0254661. doi:10.1371/journal.pone.0254661
23. 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. doi:10.1136/bmjopen-2020-037719
24. 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 Epidemiology*. 2018;57:148–157. doi:10.1016/j.canep.2018.08.009
25. Margulis AV, Fortuny J, Kaye JA, et al. Validation of cancer cases using primary care, Cancer Registry, and hospitalization data in the United Kingdom. *Epidemiology*. 2018;29(2):308–313. doi:10.1097/EDE.0000000000000786
26. Di Girolamo C, Walters S, Benitez Majano S, et al. Characteristics of patients with missing information on stage: a population-based study of patients diagnosed with colon, lung or breast cancer in England 2013. *BMC Cancer*. 2018;18(1):492. doi:10.1186/s12885-018-4417-3
27. Ranopa M, Douglas I, van Staa T, et al. The identification of incident cancers in UK primary care databases: a systematic review. *Pharmacoepidemiol Drug Safety*. 2015;24(1):11–18. doi:10.1002/pds.3729