






Presence of Breast Cancer Information Recorded in United Kingdom Primary Care Databases: Comparison of CPRD Aurum and CPRD GOLD (Companion Paper I)

Katrina Wilcox Hagberg ¹, Catherine Vasilakis-Scaramozza ², Rebecca Persson ¹, David Neasham², George Kafatos ², Susan Jick ^{1,3}

¹Epidemiology, Boston Collaborative Drug Surveillance Program, Lexington, MA, USA; ²Center for Observational Research, Amgen Ltd, Uxbridge, UK; ³Epidemiology, 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 presence of data elements related to diagnosis and treatment of malignant breast cancer in CPRD Aurum compared to those in the previously validated CPRD GOLD.

Methods: Females in CPRD Aurum or GOLD with a first-time code for malignant breast cancer, mastectomy, or ≥ 1 prescription for tamoxifen or aromatase inhibitors (2004–2019) were selected. We compared the presence of the codes for breast cancer diagnosis, surgeries (mastectomy, lumpectomy), tamoxifen and aromatase inhibitor prescriptions, radiation, chemotherapy, and supporting clinical codes (suspected breast cancer, lump symptoms, biopsy, lumpectomy, cancer care, referral/visit to specialist, palliative care). Age standardized incidence rates of breast cancer diagnosis in CPRD Aurum and GOLD were calculated.

Results: There were 131,936 eligible patients in CPRD Aurum and 69,102 patients in GOLD. A similar proportion of patients in CPRD Aurum and GOLD had codes for breast cancer diagnosis, mastectomy, drug prescriptions, lump, biopsy, lumpectomy, chemotherapy, and cancer and palliative care coded in their electronic record during follow-up. However, suspected breast cancer, radiation, and referral/visits to specialists were coded more frequently in patients in CPRD Aurum compared to GOLD. Age-standardized incidence rates were similar for CPRD Aurum and GOLD.

Conclusion: Overall, there was consistency between data elements related to malignant breast cancer recorded in CPRD Aurum and GOLD, particularly for the most informative clinical details. These findings provide reassurance that breast cancer information recorded in CPRD Aurum is generally comparable to that recorded in the previously validated CPRD GOLD and support the use of CPRD Aurum for breast cancer research.

Keywords: clinical practice research datalink, CPRD Aurum, CPRD GOLD, breast cancer, validation, data quality

Introduction

The United Kingdom (UK) Clinical Practice Research Datalink (CPRD) has been providing researchers with access to electronic healthcare records for decades. CPRD General Practice OnLine Data (GOLD), formerly General Practice Research Database, is an electronic healthcare record database collected using Vision patient management software, containing data on over 15 million patients collected over the past three decades. The CPRD GOLD data has been well described and validated with over 2,400 peer-reviewed publications in the last 30 years.^{1–7} Patient numbers in CPRD GOLD have been declining as general practitioners (GPs) switch from Vision to other patient management software systems.⁸

In response to the declining use of CPRD GOLD, CPRD began offering access to another primary care electronic data source in 2018. CPRD Aurum is an electronic medical record database sourced from Egton Medical Information Systems

(EMIS[®]) patient management software, which now encompasses over 1000 contributing practices and 30 million patients to date.^{9,10} While there are similarities between CPRD Aurum and CPRD GOLD, the quality of recording in CPRD Aurum has yet to be fully assessed. Some initial published validation assessments of CPRD Aurum suggest the recording of pulmonary embolism, myocardial infarction, diabetes, hypercholesterolemia, anemia, and malignant cancers at a specified site was of high quality for research purposes, although completeness varied by condition.^{11–14}

Understanding the characteristics, strengths, and limitations of any new electronic medical data source is critical before making decisions about its suitability and use in medical research. If CPRD Aurum is used by researchers as a replacement for, or in addition to, CPRD GOLD, it is important to understand the similarities and differences of the two data sources. The objective of this study was to evaluate CPRD Aurum by evaluating the presence of data elements coded by GPs in the CPRD Aurum electronic medical records related to malignant breast cancer (diagnosis, treatments, and care) compared to those recorded in the previously validated and widely used CPRD GOLD. We also estimated and compared crude and age-standardized breast cancer incidence rates in CPRD Aurum and CPRD GOLD. These validation assessments will help apprise the research community about the quality of breast cancer information in CPRD Aurum.

Materials and Methods

Data Sources

This study was conducted in two large, longitudinal, UK population-based electronic health record databases, CPRD Aurum and CPRD GOLD. The UK National Health Service (NHS) provides universal health coverage; therefore, no segment of the population is excluded, and the age and sex distributions of these two GP data sources are representative of the UK population.^{8,9} Participating GPs contribute deidentified data, including demographic information, medical diagnoses, symptoms, referrals, and outpatient prescriptions. Key diagnoses, treatments, and follow-up information from care provided in secondary, specialists, or hospital settings are sent to the GP. This information may not be uniformly coded into the electronic record and, therefore, may not always be available for research use.

While CPRD Aurum and CPRD GOLD arise from the UK health systems, there are differences in history, patient management software, and coding systems that may result in differences in the information available for research purposes. CPRD GOLD includes participating GPs in England, Wales, Scotland, and Northern Ireland that use Vision patient management software.^{8,15} Diagnoses and other non-prescription data recorded using the Read coding system and prescriptions are coded using Gemscript. The data were established by Value Added Medical Products in 1987 before becoming General Practice Research Database (GPRD) in 1993 and Clinical Practice Research Datalink in 2012.⁸

CPRD Aurum includes participating GPs in England, with a small number from Northern Ireland, that use the EMIS patient management software platform.⁹ Diagnoses and other non-prescription data are coded using a combination of SNOMED CT (UK edition), Read Version 2 and local EMIS Web[®] software-specific codes that have been cross-mapped to a single diagnostic code dictionary by National Health Service Digital. Prescriptions are coded using the Dictionary of Medicines and Devices codes which are a subset of the SNOMED CT terminology.^{16–18} CPRD has made CPRD Aurum data available for research purposes since 2018 and includes electronic patient data collected since 1988.

Population Selection

A source population of female patients from CPRD Aurum or CPRD GOLD who had a code for malignant breast cancer, mastectomy, and/or one or more prescription for breast cancer drugs (eg tamoxifen or aromatase inhibitors) was extracted from each database ([Supplementary Table 1](#)). Eligibility date was defined as the first of these qualifying events coded in the patient's electronic record during the study period (1 January 2004–30 June 2019). Patients were followed from the start of their electronic record or 1 January 2004, whichever came later (Start Date) through 30 June 2019, end of data collection for the practice, or the end of the patient's electronic record (death, transfer out of practice), whichever came first (End Date). To restrict the cohorts to patients with incident breast cancer, we excluded patients who had any of the following events before the patient's eligibility date: 1) codes for diagnosis of breast cancer, mastectomy, or breast cancer drug prescriptions or 2) history of prior malignant cancer at any site (except non-melanoma skin cancer).

Data Element Presence and Likelihood Classification

In CPRD Aurum and CPRD GOLD, we describe and compare the presence of the codes for breast cancer diagnosis, surgeries (eg mastectomy, lumpectomy), breast cancer drug prescriptions (eg tamoxifen, aromatase inhibitors), and supporting clinical codes that are consistent with the diagnosis and treatment of breast cancer (eg suspected breast cancer, lump symptoms, biopsy, lumpectomy, radiation, chemotherapy, cancer care, referral to specialist, palliative care). Based on the presence and relative timing of these data elements coded in CPRD Aurum and, separately, CPRD GOLD, we classified the likelihood that the patient was a true breast cancer case as likely, possible, or unsupported (Table 1).

Statistical Analyses

In the CPRD Aurum and CPRD GOLD cohorts, we describe the characteristics of patients at the eligibility date, including age, calendar year, region, and record length. We report the proportions of patients with codes for breast cancer diagnoses, surgeries, drug treatments, and supporting clinical codes recorded during follow-up. We also report the likelihood classification, overall and stratified by eligibility year and age at eligibility date.

Among patients with a coded breast cancer diagnosis, we estimated crude incidence rates of breast cancer in CPRD Aurum and CPRD GOLD overall and stratified by time period (2004–2009, 2010–2014, 2015–2018). This analysis was truncated at 2018, the last full year of available data (study period ended 30 June 2019). To compare CPRD Aurum and CPRD GOLD, incidence rates were standardized using The Office for National Statistics 2018 age-standardized data.¹⁹ Analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA) and Stata version 15.1 (StataCorp, College Station, Texas, USA).

Ethical Review and Copyright

This study is based 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. 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

There were 131,936 patients in CPRD Aurum and 69,102 patients in CPRD GOLD who qualified for analysis (Supplementary Figure 1). The mean length of follow-up was similar for patients in CPRD Aurum (11.5 ± 4.7 years)

Table 1 Breast Cancer Likelihood Classification Definition

Classification	Level	Definition
Likely	1	Breast cancer diagnosis, plus ≥ 1 treatments recorded within 365 days before or after the first breast cancer diagnosis: Breast cancer surgery, Breast cancer drug prescription, radiation, and/or chemotherapy
Possible	2	Breast cancer diagnosis plus ≥ 1 supporting clinical code recorded within 365 days before or after the first breast cancer diagnosis
	3	No breast cancer diagnosis, but patient has records for ≥ 1 breast cancer treatments: breast cancer surgery, breast cancer drug prescription, radiation, and/or chemotherapy
Unsupported	4	Breast cancer diagnosis, plus ≥ 1 of the following recorded more than 365 days before or after the first breast cancer diagnosis date: breast cancer surgery, breast cancer drug prescription, radiation, chemotherapy, or supporting clinical code
	5	Breast cancer diagnosis with no record of any breast cancer surgery, breast cancer drug prescription, radiation, chemotherapy, or supporting clinical code

Notes: Breast cancer surgery = mastectomy, lumpectomy. Breast cancer drug prescription = tamoxifen, anastrozole, letrozole, exemestane. Supporting clinical code = suspected breast cancer, lump symptom, breast biopsy, cancer care, oncology or breast clinic office visit or referral, palliative care.

and CPRD GOLD (10.7 ± 4.5 years). The proportion of patients who had a breast cancer diagnosis, mastectomy, and breast cancer drug prescription were similar in CPRD Aurum and CPRD GOLD, and breast cancer diagnosis was the first eligibility event recorded for most patients in both data sources (Table 2). In CPRD Aurum, the proportion of eligible patients was stable across the calendar periods, whereas in CPRD GOLD the proportion of eligible patients declined in the most recent time period, reflecting the overall decrease in practices contributing to CPRD GOLD. Patients in CPRD Aurum had more breast cancer diagnosis codes per patient (mean 5.8 ± 11.9) and breast cancer drug prescriptions (mean 39.6 ± 35.7) compared to patients in CPRD GOLD (breast cancer diagnosis mean 1.3 ± 1.0 , breast cancer drug prescriptions mean 31.9 ± 28.8); however, the number of mastectomy codes was similar for both data sources (Table 3). A similar proportion of patients in CPRD Aurum and CPRD GOLD had codes for breast lump symptoms, biopsy, lumpectomy, chemotherapy, cancer care, and palliative care coded in their electronic record during follow-up.

Table 2 Characteristics of the Eligible Population and Presence of Breast Cancer Diagnoses, Treatments and Other Supporting Clinical Codes, by Data Source

Characteristics	CPRD Aurum N=131,936 (%)	CPRD GOLD N=69,102 (%)
Length of follow-up (Start Date to End Date) (years)		
Mean \pm st. dev.	11.5 \pm 4.7	10.7 \pm 4.5
Median (IQR)	13.9 (7.9–15.5)	11.5 (7.4–15.5)
Region		
England		
North East	4553 (3.5)	680 (1.0)
North West	24,387 (18.5)	5857 (8.5)
Yorkshire and the Humber	4649 (3.5)	1231 (1.8)
East Midlands	3017 (2.3)	1188 (1.7)
West Midlands	22,348 (16.9)	4948 (7.2)
East of England	6798 (5.2)	3671 (5.3)
South West	16,912 (12.8)	4309 (6.2)
South Central	16,521 (12.5)	5632 (8.2)
London	19,429 (14.7)	5903 (8.5)
South East Coast	12,813 (9.7)	6633 (9.6)
Northern Ireland	499 (0.4)	3250 (4.7)
Scotland	0 (0.0)	15,160 (21.9)
Wales	0 (0.0)	10,640 (15.4)
Event recorded on Eligibility Date		
Breast cancer diagnosis only	112,885 (85.6)	58,812 (85.1)
Mastectomy only	5267 (4.0)	2467 (3.6)
Breast cancer drug prescription only	10,292 (7.8)	5666 (8.2)
Multiple categories of events coded on Eligibility Date	3492 (2.7)	2157 (3.1)

(Continued)

Table 2 (Continued).

Characteristics	CPRD Aurum N=131,936 (%)	CPRD GOLD N=69,102 (%)
Age (years) at Eligibility Date		
<30	1069 (0.8)	534 (0.8)
30–39	6536 (5.0)	3157 (4.6)
40–49	21,426 (16.2)	10,752 (15.6)
50–59	31,431 (23.8)	16,592 (24.0)
60–69	31,942 (24.2)	17,538 (25.4)
70–79	20,609 (15.6)	10,878 (15.7)
≥80	18,923 (14.3)	9651 (14.0)
Mean ± st. dev. Median (IQR range)	62.3 ± 14.9 61.8 (51.0–72.8)	62.5 ± 14.7 62.0 (51.4–72.8)
Year of first eligibility event during study period		
2004–2009	45,877 (34.8)	29,826 (43.2)
2010–2014	43,378 (32.9)	24,321 (35.2)
2015–2019	42,681 (32.4)	14,955 (21.6)

Abbreviations: IQR, Interquartile range; st.dev, Standard Deviation.

Table 3 Recording of Breast Cancer Diagnosis, Treatments, and Supporting Clinical Codes During Follow-Up in CPRD Aurum and CPRD GOLD

Breast Cancer Data Element ^a	CPRD Aurum N=131,936 (%)	CPRD GOLD N=69,102 (%)
Breast cancer diagnosis Mean number codes ± st. dev. Median (IQR)	120,084 (91.0) 5.8 ± 11.9 2 (1–5)	62,877 (91.0) 1.3 ± 1.0 1 (1–1)
Mastectomy Mean number codes ± st. dev. Median (IQR)	38,459 (29.2) 1.6 ± 2.2 1 (1–2)	19,853 (28.7) 1.1 ± 0.4 1 (1–1)
Lumpectomy Mean number codes ± st. dev. Median (IQR)	43,581 (33.0) 1.4 ± 1.7 1 (1–1)	20,001 (28.9) 1.1 ± 0.4 1 (1–1)
Breast cancer drug prescription Mean number codes ± st. dev. Median (IQR)	97,130 (73.6) 39.6 ± 35.7 32 (17–55)	51,272 (74.2) 31.9 ± 28.8 26 (11–44)
Radiation Mean number codes ± st. dev. Median (IQR)	26,044 (19.7) 1.3 ± 0.7 1 (1–1)	7967 (11.5) 1.1 ± 0.4 1 (1–1)

(Continued)

Table 3 (Continued).

Breast Cancer Data Element ^a	CPRD Aurum N=131,936 (%)	CPRD GOLD N=69,102 (%)
Chemotherapy Mean number codes ± st. dev. Median (IQR)	18,496 (14.0) 2.2 ± 2.9 1 (1–2)	9301 (13.5) 1.8 ± 2.0 1 (1–2)
Suspected breast cancer Mean number codes ± st. dev. Median (IQR)	24,800 (18.8) 1.3 ± 1.3 1 (1–1)	4201 (6.1) 1.1 ± 0.3 1 (1–1)
Lump symptom Mean number codes ± st. dev. Median (IQR)	58,434 (44.3) 1.6 ± 2.2 1 (1–2)	30,799 (44.6) 1.4 ± 0.9 1 (1–2)
Breast cancer biopsy Mean number codes ± st. dev. Median (IQR)	17,117 (13.0) 1.2 ± 0.9 1 (1–1)	8913 (12.9) 1.1 ± 0.4 1 (1–1)
Cancer care Mean number codes ± st. dev. Median (IQR)	109,088 (82.7) 2.1 ± 2.8 1 (1–2)	54,670 (79.1) 2.1 ± 2.9 1 (1–2)
Referral or visit to specialist Mean number codes ± st. dev. Median (IQR)	107,368 (81.4) 9.7 ± 9.5 7 (3–13)	47,410 (68.6) 9.1 ± 8.8 7 (3–12)
Palliative care Mean number codes ± st. dev. Median (IQR)	14,791 (11.2) 3.7 ± 5.7 2 (1–4)	6876 (10.0) 3.2 ± 4.6 1 (1–3)

Note: ^aNumber of codes restricted to 1 per date among patients with at least one code.
Abbreviations: IQR, Interquartile range; st.dev, Standard Deviation.

However, suspected breast cancer, radiation, and referral or visits to specialists were coded more frequently in patients in CPRD Aurum compared to CPRD GOLD (Table 3).

Table 4 presents the likelihood classifications, overall and stratified by calendar time, and age at eligibility date. The proportion of patients classified as likely, possible, and unsupported breast cancer cases was similar for CPRD Aurum (81.4%, 17.2%, and 1.4%, respectively) and CPRD GOLD (81.1%, 17.0%, and 1.9%, respectively) and was similar when stratified by calendar time. When stratified by age group, the likelihood classifications were similarly distributed for women aged 40 years and older. However, among women aged <30 and 30–39 years, there was a materially higher proportion of females classified as possible breast cancer cases (level 3) based on presence of codes for surgery, prescriptions, radiation and/or chemotherapy

Table 4 Breast Cancer Likelihood Classification, by Data Source, Stratified by Calendar Time and Age at Eligibility Date

CPRD Aurum					
Likelihood Classification	Likely	Possible			Unsupported
Level	1 N=107,434 (row %)	2 N=10,759 (row %)	3 N=11,852 (row %)	4 N=949 (row %)	5 N=942 (row %)
Overall	107,434 (81.4)	10,759 (8.2)	11,852 (9.0)	949 (0.7)	942 (0.7)

(Continued)

Table 4 (Continued).

CPRD Aurum					
Likelihood Classification	Likely	Possible		Unsupported	
Level	1 N=107,434 (row %)	2 N=10,759 (row %)	3 N=11,852 (row %)	4 N=949 (row %)	5 N=942 (row %)
Stratified by Eligibility Year					
2004–2009	37,864 (82.5)	3,419 (7.5)	3,713 (8.1)	431 (0.9)	450 (1.0)
2010–2014	36,135 (83.3)	3,181 (7.3)	3,564 (8.2)	248 (0.6)	250 (0.6)
2015–2019	33,435 (78.3)	4,159 (9.7)	4,575 (10.7)	270 (0.6)	242 (0.6)
Stratified by Age (years) at Eligibility Date					
<30	481 (45.0)	120 (11.2)	429 (40.1)	5 (0.5)	34 (3.2)
30–39	4,614 (70.6)	704 (10.8)	1,114 (17.0)	52 (0.8)	52 (0.8)
40–49	17,440 (81.4)	1,883 (8.8)	1,871 (8.7)	160 (0.8)	72 (0.3)
50–59	26,261 (83.6)	2,672 (8.5)	2,119 (6.7)	251 (0.8)	128 (0.4)
60–69	27,045 (84.7)	2,436 (7.6)	2,056 (6.4)	248 (0.8)	157 (0.5)
70–79	17,309 (84.0)	1,434 (7.0)	1,581 (7.7)	135 (0.7)	150 (0.7)
≥80	14,284 (75.5)	1,510 (8.0)	2,682 (14.2)	98 (0.5)	349 (1.8)
CPRD GOLD					
Likelihood Classification	Likely	Possible		Unsupported	
Level	1 N=55,949 (row %)	2 N=5,666 (row %)	3 N=6,225 (row %)	4 N=440 (row %)	5 N=822 (row %)
Overall	55,949 (81.0)	5,666 (8.2)	6,225 (9.0)	440 (0.6)	822 (1.2)
Stratified by Eligibility Year					
2004–2009	24,496 (82.1)	2,215 (7.4)	2,574 (8.6)	190 (0.6)	351 (1.2)
2010–2014	20,135 (82.8)	1,911 (7.9)	1,950 (8.0)	116 (0.5)	209 (0.9)
2015–2019	11,318 (75.7)	1,540 (10.3)	1,701 (11.4)	134 (0.9)	262 (1.8)
Stratified by Age (years) at Eligibility Date					
<30	234 (43.8)	48 (9.0)	223 (3.6)	4 (0.8)	25 (4.7)
30–39	2,159 (68.4)	311 (9.9)	607 (19.2)	27 (0.9)	53 (1.7)
40–49	8,816 (82.0)	924 (8.6)	883 (8.2)	65 (0.6)	64 (0.6)
50–59	13,789 (83.1)	1,479 (8.9)	1,034 (6.3)	120 (0.7)	170 (1.0)
60–69	14,835 (84.6)	1,345 (7.7)	1,062 (6.1)	118 (0.7)	178 (1.0)
70–79	8,959 (82.4)	847 (7.8)	878 (8.1)	63 (0.6)	131 (1.2)
≥80	7,157 (74.2)	712 (7.4)	1,538 (16.0)	43 (0.5)	201 (2.1)

Table 5 Crude and Standardized Incidence Rates of Female Breast Cancer Diagnosis (per 100,000 Person-Years), by Data Source and Calendar Time

Calendar Period	CPRD Aurum			CPRD GOLD		
	2004–2009	2010–2014	2015–2018	2005–2009	2010–2014	2015–2018
Breast cancer diagnosis	42,000	39,804	34,025	27,167	22,410	12,084
Person-years	33,767,297	30,434,609	26,034,492	20,404,879	16,143,632	8,899,621
Crude Incidence Rate (95% CI)	124.4 (123.2–125.6)	130.8 (129.5–132.1)	130.7 (129.3–132.1)	133.1 (131.6–134.7)	138.8 (137.0–140.6)	135.8 (133.4–138.2)
Age-Standardized Incidence Rate ^a (95% CI)	128.9 (127.7–130.2)	135.6 (134.2–136.9)	136.8 (135.4–138.3)	133.5 (131.9–135.1)	137.6 (135.8–139.5)	133.1 (130.7–135.5)

Note: ^aStandardized using Office for National Statistics 2018 age- and sex-specific population estimates.¹⁹

Abbreviation: CI, confidence interval.

in the absence of a breast cancer diagnosis, in CPRD Aurum (level 3: 40.1% for <30 years and 17.0% for 30–39 years) compared to CPRD GOLD (level 3: 3.6% for <30 years and 9.8% for 30–39 years).

The crude incidence of breast cancer was slightly higher in CPRD GOLD for all time periods, but the age-standardized incidence rates were similar for both CPRD Aurum and CPRD GOLD (Table 5). The overall age-standardized incidence was 133.3 (95% CI 132.6–134.1) and 134.9 (95% CI 133.8–136.0) per 100,000 person-years for CPRD Aurum and CPRD GOLD, respectively.

Discussion

Overall, the recording of breast cancer diagnoses, treatments, and prescription drugs was similar for CPRD Aurum and CPRD GOLD. Slight differences were found for a few of the data elements under study, but overall, there was consistency between the two primary care databases, particularly for the most informative clinical details. In addition, the age-standardized incidence rates of breast cancer were similar between CPRD Aurum and CPRD GOLD, suggesting that breast cancer capture is quite high in CPRD Aurum. These findings provide reassurance that breast cancer information recorded in CPRD Aurum is generally comparable to that recorded in the previously validated CPRD GOLD and also support the use of CPRD Aurum for research on breast cancer.

CPRD Aurum and CPRD GOLD are both primary care medical databases managed by the CPRD. Patients with malignant breast cancer in CPRD Aurum and CPRD GOLD had a similar age distribution and presence of codes for breast cancer diagnosis, mastectomy, chemotherapy, breast lump, breast biopsy, cancer care, palliative care, and prescriptions for breast cancer drugs. There are, however, some notable differences between CPRD Aurum and CPRD GOLD data. The proportion of patients with codes for radiation, suspected breast cancer, and referrals or visits to specialists was higher in CPRD Aurum, with notable differences in presence of codes for breast cancer treatments for younger women without a malignant breast cancer diagnosis. In addition, GPs who contribute to CPRD Aurum and CPRD GOLD are differently distributed across geographic regions. The distribution will likely continue to change over time due to GP migration to different patient management software and CPRD recruitment decisions.

Our CPRD GOLD results are similar to previous breast cancer studies in the same data source with respect to age distribution, and presence of breast cancer diagnosis, treatment, symptom, and cancer care codes, which provides reassurance in the results from our comparator database. In addition, the age-standardized breast cancer incidence rates for CPRD Aurum and CPRD GOLD reported in this study are within the range reported by other UK data sources, which range from 122.5 to 169.8 per 100,000 depending on country, calendar year, and the choice of standard population.^{4,20–23} Overall, the findings of this study provide reassurance that breast cancer information, where present, captured in CPRD Aurum is of similar quality as the well described and widely used CPRD GOLD.

The NHS requires that information from hospitalizations and outpatient services be sent to the GP; therefore, we expect that GPs would be aware of malignant breast cancer diagnoses. However, GPs or their staff must code diagnoses, treatments,

and prescriptions received in hospital or in specialist clinics into the electronic record for these details to be available for use in research, which may not always be done. In addition, the diagnosis and prescription coding systems used in CPRD Aurum and CPRD GOLD are different. Text string searches were used to create all diagnostic and drug prescription code lists and to match codes between the two data sources. It is possible that code lists were more complete for one database than the other. However, the recordings of breast cancer diagnoses, surgery, and prescriptions were similar for CPRD Aurum (91.0%, 29.2%, and 73.6%, respectively) and CPRD GOLD (91.0%, 28.7%, and 74.2%, respectively). In addition, the age-standardized incidence of breast cancer was similar between the two databases. This suggests that no important codes were missed for these data elements in either data source. Most (91%) breast cancer cases in our study were identified using diagnostic codes alone, suggesting that researchers would miss approximately 9% of malignant breast cancer cases in CPRD Aurum and CPRD GOLD if relying solely on diagnostic codes, a finding that is consistent with other reports.^{21,24} Therefore, in both CPRD Aurum and CPRD GOLD, capture of cases in the absence of a malignant breast cancer diagnosis code can be improved with the inclusion of disease-specific treatments, symptoms, or cancer care codes in the case definition. Researchers may also consider using linked Hospital Episode Statistics or Cancer Registry data to improve capture of cancer cases.^{14,25,26}

The study period was 1 January 2004 through 30 June 2019; therefore, this study does not describe the recording of breast cancer diagnoses, treatments, and prescription drugs before 2004. The Quality and Outcome Framework (QOFs) in the UK was implemented in 2004, which incentivized GPs to make improvements to recording of medical information.²⁷ The quality of cancer diagnoses in CPRD GOLD prior to 2004 has been previously described.^{24,28–30} In a prior evaluation of malignant cancer diagnoses recorded in CPRD Aurum, completeness of recordings increased over the study period (1997–2017), suggesting that recording of breast cancer diagnoses before 2004 may be incomplete.¹⁴

Conclusion

This study indicates that the recording of breast cancer diagnoses, treatments, and prescription drugs was similarly high quality for CPRD Aurum and CPRD GOLD, which provides support for the use of CPRD Aurum data for research. Additional information on the quality and completeness of malignant breast cancer coding in CPRD Aurum and CPRD GOLD compared to Hospital Episodes Statistics and the Cancer Registry are described in a companion publication.²⁵ The results of this study cannot be generalized to other cancer types, specifically those that are managed in secondary or hospital settings; therefore, additional CPRD Aurum validation studies should be conducted for other cancers.

Funding

This study was funded by Amgen, Ltd.

Disclosure

Ms Katrina Wilcox Hagberg, Dr Catherine Vasilakis-Scaramozza, Ms Rebecca Persson, and Dr 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. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ*. 1991;302:766–768. doi:10.1136/bmj.302.6779.766
2. Jick SS, Kaye JA, Vasilakis-Scaramozza C, et al. Validity of the general practice research database. *Pharmacotherapy*. 2003;23:686–689. doi:10.1592/phco.23.5.686.32205
3. 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. doi:10.1111/j.1365-2125.2009.03537.x
4. 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. doi:10.3399/bjgp10X483562
5. Medicines & Healthcare Products Regulatory Agency. Clinical Practice Research Datalink [internet]. Available from: www.CPRD.com. Accessed November 30, 2023.
6. Walley T, Mantgani A. The UK General Practice Research Database. *Lancet*. 1997;350:1097–1099. doi:10.1016/S0140-6736(97)04248-7

7. CPRD bibliography. Available from: <https://cprd.com/bibliography>. Accessed November 30, 2023.
8. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44:827–836. doi:10.1093/ije/dyv098
9. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epi*. 2019;48:1740–1740g. doi:10.1093/ije/dyz034
10. CPRD Aurum May 2022 release notes. Available from: <https://cprd.com/sites/default/files/2022-05/2022-05%20CPRD%20Aurum%20Release%20Notes.pdf>. Accessed November 30, 2023.
11. 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
12. 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;13:745–775. doi:10.2147/CLEP.S319245
13. 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;29:1456–1464. doi:10.1002/pds.5135
14. 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. *Ann Cancer Epi*. 2022;6:6. doi:10.21037/ace-22-4
15. Williams T, van Staa T, Puri S, et al. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv Drug Saf*. 2012;3(2):89–99. doi:10.1177/2042098611435911
16. NHS Digital. SNOMED CT [internet]. Available from: <https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct>. Accessed November 30, 2023.
17. NHS Digital. Read codes [internet]. Available from: <https://digital.nhs.uk/services/terminology-and-classifications/read-codes>. Accessed November 30, 2023.
18. NHS Business Services Authority. Dictionary of medicines and devices (dm+d) [Internet]. Available from: <https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd>. Accessed November 30, 2023.
19. Office for National Statistics. National records of Scotland, Northern Ireland Statistics and Research Agency—population estimates 2018 data. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2020#the-uk-population-at-mid-2020>. Accessed November 30, 2023.
20. Din NU, Ukoumunne OC, Rubin G, et al. Age and gender variations in cancer diagnostic intervals in 15 cancers: analysis of data from the UK Clinical Practice Research Datalink. *PLoS One*. 2015;10(5):e0127717. doi:10.1371/journal.pone.0127717
21. 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
22. Walker S, Hyde C, Hamilton W. Risk of breast cancer in symptomatic women in primary care: a case-control study using electronic records. *Br J Gen Pract*. 2014;64(629):e788–93. doi:10.3399/bjgp14X682873
23. Dewis R, Gribbin J. *Breast Cancer: Diagnosis and Treatment: An Assessment of Need*. Cardiff (UK): National Collaborating Centre for Cancer (UK); February, 2009.
24. Kaye JA, Derby LE, Del Mar Melero-Montes M, et al. The incidence of breast cancer in the General Practice Research Database compared with national cancer registration data. *Br J Cancer*. 2000;83(11):1556–1558. doi:10.1054/bjoc.2000.1493
25. Hagberg KW, Vasilakis-Scaramozza C, Persson R, Neasham D, Kafatos G, Jick S. 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*. 2023;15:1193–1206. doi:10.2147/CLEP.S434829.
26. Jick SS, Vasilakis-Scaramozza C, Persson R, Neasham D, Kafatos G, Hagberg KW. Use of the CPRD Aurum Database: Insights Gained from New Data Quality Assessments. *Clin Epidemiol*. 2023;15:1219–1222. doi:10.2147/CLEP.S434832
27. 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. doi:10.1186/1471-2458-12-329
28. 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;57:148–157. doi:10.1016/j.canep.2018.08.009
29. 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(8):e037719. doi:10.1136/bmjopen-2020-037719
30. Trafford AM, Parisi R, Rutter MK, et al.; Global psoriasis atlas (gpa). 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

Clinical Epidemiology

Dovepress

Publish your work in this journal

Clinical Epidemiology is an international, peer-reviewed, open access, online journal focusing on disease and drug epidemiology, identification of risk factors and screening procedures to develop optimal preventative initiatives and programs. Specific topics include: diagnosis, prognosis, treatment, screening, prevention, risk factor modification, systematic reviews, risk & safety of medical interventions, epidemiology & biostatistical methods, and evaluation of guidelines, translational medicine, health policies & economic evaluations. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use.

Submit your manuscript here: <https://www.dovepress.com/clinical-epidemiology-journal>