ORIGINAL RESEARCH

Comparison of Rheumatoid Arthritis Information Recorded in UK CPRD Aurum and CPRD GOLD Databases (Companion Paper 3)

Catherine Vasilakis-Scaramozza [1], Katrina Wilcox Hagberg [1], Rebecca Persson [1], George Kafatos [2], Joe Maskell[2], David Neasham[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 report distribution of codes associated with a rheumatoid arthritis (RA) diagnosis recorded in Clinical Practice Research Datalink (CPRD) Aurum compared to the previously validated CPRD GOLD database as a critical step toward making decisions about CPRD Aurum's suitability for medical research.

Patients and Methods: We analyzed the distribution of codes for RA diagnoses, labs, and treatments in the new CPRD Aurum database, compared to the CPRD GOLD database by selecting relevant indicators of RA diagnosis, treatment, and clinical care. We included all patients in England in CPRD Aurum and CPRD GOLD with an incident diagnosis code for RA on or after 1 January 2005 and at least two years recorded data before first RA diagnosis.

Results: We found 53,083 and 18,167 patients with a new diagnosis code for RA in CPRD Aurum and CPRD GOLD, respectively. In both databases approximately 67% were female with similar mean ages at first diagnosis. There were few differences in RA-related recording patterns between the two data sources. Before first RA diagnosis, CPRD Aurum patients had more RA-specific labs and other supporting clinical codes. After diagnosis, CPRD Aurum patients had more RA diagnoses coded and more often had 10+ general RA labs than patients in CPRD GOLD. More CPRD GOLD patients had 10+ prescriptions for conventional disease-modifying antirheumatic drugs (cDMARD) compared to CPRD Aurum. Otherwise, the distribution of drugs used to treat RA was similar between databases. The standardized incidence of RA was similar between databases.

Conclusion: Overall, among patients with a diagnosis code for RA, recording of diagnoses, prescription drugs, and labs were similar between CPRD Aurum and CPRD GOLD. Slight differences were found for a few variables, but overall, we found consistency between the databases. In addition, standardized incidence of RA was similar between databases.

Keywords: Clinical Practice Research Datalink, CPRD Aurum, CPRD GOLD, validation, data quality

Introduction

The UK Clinical Practice Research Datalink (CPRD) Aurum is a medical record database made available to researchers in 2018.[1] Like CPRD GOLD, which has been well described and validated, CPRD Aurum provides de-identified data based on electronic health care records entered by general practitioners (GPs). However, understanding CPRD Aurum's data characteristics, strengths, and limitations, and comparing CPRD Aurum to the well-studied CPRD GOLD, is a critical step toward making decisions about its suitability for use in medical research.

Recent studies of CPRD Aurum suggest that the quality and completeness of the data are high, but that additional validation studies were needed to describe the suitability of the data for use in medical research on various exposures and outcomes.[2–5] Researchers have begun using CPRD Aurum for studies, some of which combine CPRD Aurum data and CPRD GOLD data.[6–9] However, the only other published validation studies comparing CPRD Aurum and CPRD GOLD were a study of antibiotic prescribing,[10] a feasibility study to identify patients with chronic obstructive pulmonary disease,[11] and a study of breast cancer

1207

published in this issue of *Clinical Epidemiology*.[12] CPRD GOLD has been successfully used in many studies of rheumatoid arthritis (RA) and other chronic outcomes,[13–20] but assessment of data recording for RA in CPRD Aurum is limited to our previous short report that found that codes expected for monitoring and treatment of RA are routinely recorded in CPRD Aurum, without comparison to an external data source.[21]

We expanded on this by further evaluating RA diagnoses and treatments in the new CPRD Aurum database to data in the previously validated CPRD GOLD database. Our objectives were to compare the distribution of codes for diagnoses, tests, and treatments consistent with a diagnosis of RA in CPRD Aurum and CPRD GOLD. We also estimate and compare age- and sex-standardized incidence rates of RA in CPRD Aurum and CPRD GOLD.

Materials and Methods
Data Sources
CPRD Aurum and CPRD GOLD are two large, longitudinal, population-based electronic health record databases based on GP patient management computing systems in the UK that have been described extensively elsewhere[1,22] including in our study of breast cancer published in this issue of *Clinical Epidemiology*.[12] Briefly, CPRD Aurum and CPRD GOLD both arise from the UK health services and are based on de-identified data recorded by GPs. The databases differ, however, in that they are sourced from two different patient management software systems (EMIS and Vision), and in their geographic coverage, coding systems, and population size.[1,22] The August 2020 CPRD Aurum data build and the January 2020 CPRD GOLD flat file were used for this study.

Population Selection
This was a descriptive study of RA diagnoses, labs, and treatments recorded in CPRD Aurum compared to CPRD GOLD. Subjects in the study were patients from CPRD Aurum and CPRD GOLD practices in England who had a first diagnosis code for RA recorded between 2005 and 2019 in the absence of any indication that such a diagnosis was made in the past, and who had at least two years of data in their electronic record before their first RA diagnosis date and after registration (to ensure that the RA diagnosis was incident). These patients were followed through 30 June 2019 (end of data collection) or the end of the patient record (death, transfer out of practice), whichever came first. The start and end of each patient's active CPRD electronic record was estimated using available registration, prescription, and clinical data.[5]

Statistical Analyses
We present the distribution of patient characteristics and recorded codes associated with RA for each database. No formal statistical testing was conducted. We describe age and sex of patients at first RA diagnosis (index date) (mean age in years ± SD) and time in the database (median and interquartile range), stratified by time period (2005–2009, 2010–2014, 2015–2019) and data source. The median patient-time in the database before and after the date of first RA diagnosis differed between CPRD Aurum and CPRD GOLD, and as a result we restricted our comparisons to one year before or after the index date so that any potential differences were not a result of disparities in available time in each database. For each database we report the proportion of patients who had codes for labs that are typically used to diagnose and monitor RA, treatments, and relevant clinical care codes recorded in the one year before and after first RA diagnosis, by time period and data source (CPRD Aurum or CPRD GOLD), and also the number of times the codes occurred in the patient record during the study period. We evaluated the presence of lab tests, but not the results of the tests. RA-specific labs, used to diagnose and monitor RA, were rheumatoid factor and anti-CCP antibody. General RA labs, used to diagnose and monitor RA as well as other autoimmune disorders, were antinuclear antibody (ANA), C-reactive protein (CRP), human leukocyte antigen (HLA) and erythrocyte sedimentation rate. Analgesics included NSAIDs and opioids. Supporting clinical codes included rheumatology referrals and visits, drug monitoring, and joint pain codes. We selected patient treatment, diagnosis, and clinical care codes for comparison based on the NICE guidelines as a starting point of reference for codes we would expect to find in the medical records of patients with RA.[23,24] These codes are provided in Supplemental Table 1.

We also report the crude and age- and sex-standardized incidence of RA (defined as presence of one or more RA diagnostic codes) per 100,000 person-years by time period and data source. For the denominator we accumulated person-time for all patients in each data source starting two years after registration date or 1 January 2005, whichever was later, to be consistent

with RA case eligibility requirements. We stopped accumulating person-time on the day a patient transferred out (including death) or the last data collection date for the practice, whichever was first. To compare CPRD Aurum and CPRD GOLD, incidence rates were standardized using The Office for National Statistics 2018 age- and sex-specific population estimates.[25] All analyses were performed using Stata version 15.1 (StataCorp, College Station, TX, USA).

Ethical Review Statement

This study is based on data from the CPRD 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 Research Data Governance Committee (RDG) for Medicines and Healthcare Products Regulatory Agency (protocol no: 20_000064), and the protocol was made available to the journal reviewers upon request.

Results

Within practices in England, we found 53,083 patients with a new diagnosis code for RA in CPRD Aurum and 18,167 patients in CPRD GOLD, with at least two years of recorded data before first RA diagnosis, between 2005 and 2019. The number of patients across time periods in CPRD Aurum was reasonably consistent, but the number of patients in CPRD GOLD dropped considerably, especially in the most recent time period, reflecting the overall decrease in general practices contributing to CPRD GOLD.

In both CPRD Aurum and CPRD GOLD, approximately 67% of patients diagnosed with RA were female. This finding was consistent over time and between CPRD Aurum and CPRD GOLD (Table 1). The mean age of patients at

Table 1 Characteristics of RA Cases, by Data Source at Time of First RA Diagnosis

	CPRD Aurum			CPRD GOLD		
	2005–2009	**2010–2014**	**2015–2019**	**2005–2009**	**2010–2014**	**2015–2019**
	N=16,489 (%)	**N=18,044 (%)**	**N=18,550 (%)**	**N=8,451 (%)**	**N=7,085 (%)**	**N=2,631 (%)**
Sex						
Female	11,197 (68%)	12,106 (67%)	12,431 (67%)	5782 (68%)	4732 (67%)	1748 (66%)
Male	5292 (32%)	5938 (33%)	6119 (33%)	2669 (32%)	2353 (33%)	883 (34%)
Age (years) at first RA diagnosis						
<30	569 (3%)	718 (4%)	731 (4%)	477 (6%)	369 (5%)	135 (5%)
30–39	1204 (7%)	1369 (8%)	1330 (7%)	664 (8%)	572 (8%)	217 (8%)
40–49	2449 (15%)	2738 (15%)	2650 (14%)	1240 (15%)	1052 (15%)	401 (15%)
50–59	3582 (22%)	3907 (22%)	4218 (23%)	1788 (21%)	1543 (22%)	593 (23%)
60–69	3794 (23%)	4148 (23%)	4142 (22%)	1930 (23%)	1608 (23%)	562 (21%)
70–79	3244 (20%)	3573 (20%)	3771 (20%)	1561 (18%)	1279 (18%)	508 (19%)
≥80	1647 (10%)	1591 (9%)	1708 (9%)	791 (9%)	662 (9%)	215 (8%)
Mean ± SD	60 (16)	59 (16)	60 (16)	59 (17)	58 (16)	58 (16)
Time in database in years, median (IQR)						
Total time in database	20 (15–25)	20 (14–25)	19 (11–25)	19 (14–23)	19 (13–24)	20 (12–26)
Time in database before index date	11 (7–15)	14 (8–19)	16 (9–23)	13 (8–16)	15 (9–20)	18 (11–24)
Time in database after index date	10 (5–12)	6 (5–8)	2 (1–3)	6 (4–9)	3 (2–5)	1 (1–3)

first diagnosis code for RA was 59.5 years (SD 15.6) in CPRD Aurum and 58.5 years (SD 16.4) in CPRD GOLD and was also consistent across time periods (Table 1). Although the total median patient-time in the database was similar between the two databases (median 19.5 years [IQR 13.5–24.8] for CPRD Aurum and median 18.9 [IQR 13.3–23.6] for CPRD GOLD), the median patient-time in the database before and after the date of first RA diagnosis differed between the two sources (Table 1). Within each time period, patients in CPRD GOLD had more time in the database before first RA diagnosis compared to CPRD Aurum, and patients in CPRD Aurum had more time in the database after first RA diagnosis compared to CPRD GOLD (Table 1). As a result, further analyses were restricted to one year before or after the date of first RA diagnosis so that any differences in recorded diagnoses, labs or drugs were not a result of differences in available time in each database. Most patients had more than one year of data recorded after their first RA diagnosis, with only 13% of patients in CPRD Aurum and 14% of patients in CPRD GOLD having less than one year of data after first RA diagnosis.

Table 2 and Table 3 provide information on the distribution of recorded events for diagnoses, labs, and prescriptions within a year before first RA diagnosis (Table 2), and within a year after first RA diagnosis (Table 3), by time period, for CPRD Aurum and CPRD GOLD. Biologic prescriptions are not included in the table because few patients (<1%) in either database had a biologic recorded by the GP. In the year before and the year after first RA diagnosis the distribution of diagnoses, labs, and prescriptions was similar between CPRD Aurum and CPRD GOLD, with a few minor exceptions. Before first RA diagnosis we found slight differences between the databases in the distribution of supporting clinical codes (rheumatology referrals and visits, drug monitoring, and joint pain codes) (Table 2). The recording of supporting clinical codes increased over time in both databases both before and after first RA diagnosis, but these clinical codes were recorded in CPRD Aurum more often, particularly in the time period 2015–2019 (Table 2 and Table 3). Patients in CPRD GOLD more often had only one RA diagnosis recorded compared to patients in CPRD Aurum who more often had two or more diagnoses recorded (Table 3). The number of patients with general RA labs after first diagnosis was

Table 2 Counts of Lab, Treatment, and Supporting Clinical Codes Within One Year Before First RA Diagnosis, by Data Source and Time Period (of First RA Diagnosis)

	CPRD Aurum			CPRD GOLD		
	2005–2009	**2010–2014**	**2015–2019**	**2005–2009**	**2010–2014**	**2015–2019**
	N=16,489 (%)	**N=18,044 (%)**	**N=18,550 (%)**	**N=8451 (%)**	**N=7085 (%)**	**N=2631 (%)**
Number of RA-specific lab codes[a]						
0	8653 (52%)	8387 (46%)	8980 (48%)	4200 (50%)	3181 (45%)	1259 (48%)
1	4908 (30%)	6526 (36%)	5789 (31%)	3442 (41%)	2914 (41%)	958 (36%)
2–4	2839 (17%)	3046 (17%)	3716 (20%)	802 (9%)	988 (14%)	411 (16%)
≥5	89 (1%)	85 (0%)	65 (0%)	7 (0%)	2 (0%)	3 (0%)
Median (IQR)	0 (0–1)	1 (0–1)	1 (0–1)	1 (0–1)	1 (0–1)	1 (0–1)
Number of RA (general) lab codes[a]						
0	5001 (30%)	3909 (22%)	3668 (20%)	2511 (30%)	1525 (22%)	620 (24%)
1	3028 (18%)	2654 (15%)	2695 (15%)	1535 (18%)	1011 (14%)	380 (14%)
2–4	6031 (37%)	8042 (45%)	8696 (47%)	3151 (37%)	3224 (46%)	1178 (45%)
5–9	1688 (10%)	2270 (13%)	2375 (13%)	870 (10%)	917 (13%)	323 (12%)
≥10	741 (4%)	1169 (6%)	1116 (6%)	384 (5%)	408 (6%)	130 (5%)
Median (IQR)	2 (0–3)	2 (1–4)	2 (1–4)	2 (0–3)	2 (1–4)	2 (1–4)

(*Continued*)

Table 2 (Continued).

	CPRD Aurum			CPRD GOLD		
	2005–2009	**2010–2014**	**2015–2019**	**2005–2009**	**2010–2014**	**2015–2019**
	N=16,489 (%)	**N=18,044 (%)**	**N=18,550 (%)**	**N=8451 (%)**	**N=7085 (%)**	**N=2631 (%)**
Number of cDMARD prescriptions						
0	13,282 (81%)	14,127 (78%)	14,436 (78%)	6797 (80%)	5553 (78%)	2039 (77%)
I	672 (4%)	771 (4%)	797 (4%)	350 (4%)	290 (4%)	118 (4%)
2–4	861 (5%)	1070 (6%)	1017 (5%)	457 (5%)	450 (6%)	141 (5%)
5–9	891 (5%)	968 (5%)	1029 (6%)	514 (6%)	437 (6%)	180 (7%)
≥10	783 (5%)	1108 (6%)	1271 (7%)	333 (4%)	355 (5%)	153 (6%)
Median (IQR)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)
Number of analgesic[b] prescriptions						
0	3381 (21%)	4053 (22%)	5293 (29%)	1850 (22%)	1703 (24%)	725 (28%)
I	2075 (13%)	2546 (14%)	2765 (15%)	1052 (12%)	992 (14%)	398 (15%)
2–4	4218 (26%)	4799 (27%)	4592 (25%)	2225 (26%)	1913 (27%)	656 (25%)
5–9	3556 (22%)	3436 (19%)	2972 (16%)	1724 (20%)	1319 (19%)	431 (16%)
≥10	3259 (20%)	3210 (18%)	2928 (16%)	1600 (19%)	1158 (16%)	421 (16%)
Median (IQR)	3 (1–8)	3 (1–7)	2 (0–6)	3 (1–8)	3 (1–7)	2 (0–6)
Number of NSAID prescriptions						
0	5874 (36%)	7492 (42%)	9218 (50%)	3001 (36%)	2915 (41%)	1222 (46%)
I	2595 (16%)	3059 (17%)	3202 (17%)	1370 (16%)	1226 (17%)	463 (18%)
2–4	4423 (27%)	4583 (25%)	3958 (21%)	2306 (27%)	1836 (26%)	602 (23%)
5–9	2553 (15%)	2126 (12%)	1543 (8%)	1269 (15%)	811 (11%)	252 (10%)
≥10	1044 (6%)	784 (4%)	629 (3%)	505 (6%)	297 (4%)	92 (3%)
Median (IQR)	1 (0–4)	1 (0–3)	1 (0–2)	1 (0–4)	1 (0–3)	1 (0–3)
Number of opioid prescriptions						
0	7732 (47%)	8265 (46%)	9139 (49%)	4204 (50%)	3458 (49%)	1318 (50%)
I	2464 (15%)	2874 (16%)	2839 (15%)	1180 (14%)	1088 (15%)	390 (15%)
2–4	2885 (17%)	3179 (18%)	2910 (16%)	1415 (17%)	1204 (17%)	408 (16%)
5–9	1776 (11%)	1855 (10%)	1716 (9%)	879 (10%)	677 (10%)	262 (10%)
≥10	1632 (10%)	1871 (10%)	1946 (10%)	773 (9%)	658 (9%)	253 (10%)
Median (IQR)	1 (0–3)	1 (0–3)	1 (0–3)	1 (0–3)	1 (0–3)	0 (0–3)

(*Continued*)

Table 2 (Continued).

	CPRD Aurum			CPRD GOLD		
	2005–2009	**2010–2014**	**2015–2019**	**2005–2009**	**2010–2014**	**2015–2019**
	N=16,489 (%)	**N=18,044 (%)**	**N=18,550 (%)**	**N=8451 (%)**	**N=7085 (%)**	**N=2631 (%)**
Number of glucocorticoid prescriptions						
0	12,328 (75%)	12,710 (70%)	12,828 (69%)	6110 (72%)	4888 (69%)	1832 (70%)
1	1328 (8%)	1772 (10%)	1971 (11%)	829 (10%)	772 (11%)	290 (11%)
2–4	1329 (8%)	1860 (10%)	2010 (11%)	792 (9%)	802 (11%)	278 (11%)
5–9	859 (5%)	943 (5%)	1013 (5%)	477 (6%)	413 (6%)	156 (6%)
≥10	645 (4%)	759 (4%)	728 (4%)	243 (3%)	210 (3%)	75 (3%)
Median (IQR)	0 (0–1)	0 (0–1)	0 (0–1)	0 (0–1)	0 (0–1)	0 (0–1)
Number of supporting clinical codes[c]						
0	9798 (59%)	6120 (34%)	4012 (22%)	5189 (61%)	2496 (35%)	844 (32%)
1	4441 (27%)	5404 (30%)	5232 (28%)	1623 (19%)	1668 (24%)	641 (24%)
2–4	1878 (11%)	5102 (28%)	7260 (39%)	1382 (16%)	2372 (33%)	934 (35%)
5–9	310 (2%)	1219 (7%)	1795 (10%)	249 (3%)	494 (7%)	192 (7%)
≥10	62 (0%)	199 (1%)	251 (1%)	8 (0%)	55 (1%)	20 (1%)
Median (IQR)	0 (0–1)	1 (0–2)	2 (1–3)	0 (0–1)	1 (0–2)	1 (0–3)

Notes: [a]RA-specific labs included are rheumatoid factor and anti-CCP antibody. General RA labs included are antinuclear antibody (ANA), C-reactive protein (CRP), human leukocyte antigen (HLA) and erythrocyte sedimentation rate. [b]Analgesics include NSAIDs and opioids. [c]Supporting clinical codes include rheumatology referrals and visits, drug monitoring, and joint pain codes.

Table 3 Counts of Lab, Treatment, and Supporting Clinical Codes Within One Year After First RA Diagnosis, by Data Source and Time Period (of First RA Diagnosis)

	CPRD Aurum			CPRD GOLD		
	2005–2009	**2010–2014**	**2015–2019**	**2005–2009**	**2010–2014**	**2015–2019**
	N=16,489 (%)	**N=18,044 (%)**	**N=18,550 (%)**	**N=8,451 (%)**	**N=7,085 (%)**	**N=2,631 (%)**
Number of RA diagnosis codes						
0	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
1	8196 (50%)	8161 (45%)	8798 (47%)	5979 (71%)	5258 (74%)	2084 (79%)
2–4	5221 (32%)	6195 (34%)	6816 (37%)	2218 (26%)	1678 (24%)	521 (20%)
5–9	1948 (12%)	2369 (13%)	2090 (11%)	226 (3%)	121 (2%)	25 (1%)
≥10	1124 (7%)	1319 (7%)	846 (5%)	28 (0%)	28 (0%)	1 (0%)
Median (IQR)	2 (1–3)	2 (1–4)	2 (1–3)	1 (1–2)	1 (1–2)	1 (1–1)

(Continued)

Table 3 (Continued).

	CPRD Aurum			CPRD GOLD		
	2005–2009	**2010–2014**	**2015–2019**	**2005–2009**	**2010–2014**	**2015–2019**
	N=16,489 (%)	**N=18,044 (%)**	**N=18,550 (%)**	**N=8,451 (%)**	**N=7,085 (%)**	**N=2,631 (%)**
Number of RA-specific lab codes[a]						
0	14,055 (85%)	16,127 (89%)	17,205 (93%)	7291 (86%)	6318 (89%)	2399 (91%)
1	1580 (10%)	1353 (7%)	964 (5%)	952 (11%)	563 (8%)	172 (7%)
2–4	800 (5%)	538 (3%)	362 (2%)	195 (2%)	200 (3%)	58 (2%)
≥5	54 (0%)	26 (0%)	19 (0%)	13 (0%)	4 (0%)	2 (0%)
Median (IQR)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)
Number of RA (general) lab codes[a]						
0	5657 (34%)	4681 (26%)	5622 (30%)	3117 (37%)	2036 (29%)	917 (35%)
1	1665 (10%)	1509 (8%)	1660 (9%)	884 (10%)	629 (9%)	227 (9%)
2–4	3291 (20%)	3719 (21%)	4202 (23%)	1735 (21%)	1534 (22%)	649 (25%)
5–9	1976 (12%)	2572 (14%)	2816 (15%)	1033 (12%)	994 (14%)	365 (14%)
≥10	3900 (24%)	5563 (31%)	4250 (23%)	1682 (20%)	1892 (27%)	473 (18%)
Median (IQR)	2 (0–9)	4 (0–12)	2 (0–8)	2 (0–7)	3 (0–10)	2 (0–6)
Number of cDMARD prescriptions						
0	5418 (33%)	4318 (24%)	5162 (28%)	3192 (38%)	1995 (28%)	828 (31%)
1	657 (4%)	731 (4%)	1251 (7%)	383 (5%)	295 (4%)	163 (6%)
2–4	1802 (11%)	1913 (11%)	2481 (13%)	1044 (12%)	891 (13%)	398 (15%)
5–9	3662 (22%)	3728 (21%)	3674 (20%)	1933 (23%)	1598 (23%)	524 (20%)
≥10	4950 (30%)	7354 (41%)	5982 (32%)	1899 (22%)	2306 (33%)	718 (27%)
Median (IQR)	5 (0–11)	7 (1–13)	5 (0–11)	3 (0–9)	6 (0–11)	4 (0–10)
Number of analgesic[b] prescriptions						
0	4089 (25%)	6143 (34%)	8358 (45%)	2332 (28%)	2603 (37%)	1138 (43%)
1	1784 (11%)	2084 (12%)	2277 (12%)	898 (11%)	802 (11%)	331 (13%)
2–4	2950 (18%)	3155 (17%)	2809 (15%)	1558 (18%)	1232 (17%)	457 (17%)
5–9	3321 (20%)	2804 (16%)	2312 (12%)	1613 (19%)	1120 (16%)	322 (12%)
≥10	4345 (26%)	3858 (21%)	2794 (15%)	2050 (24%)	1328 (19%)	383 (15%)
Median (IQR)	4 (1–10)	2 (0–8)	1 (0–5)	3 (0–9)	2 (0–7)	1 (0–5)

(*Continued*)

Table 3 (Continued).

	CPRD Aurum			CPRD GOLD		
	2005–2009	**2010–2014**	**2015–2019**	**2005–2009**	**2010–2014**	**2015–2019**
	N=16,489 (%)	**N=18,044 (%)**	**N=18,550 (%)**	**N=8,451 (%)**	**N=7,085 (%)**	**N=2,631 (%)**
Number of NSAID prescriptions						
0	7357 (45%)	10377 (58%)	12737 (69%)	3988 (47%)	4109 (58%)	1728 (66%)
I	1882 (11%)	2025 (11%)	1855 (10%)	977 (12%)	771 (11%)	288 (11%)
2–4	2721 (17%)	2413 (13%)	1913 (10%)	1353 (16%)	998 (14%)	314 (12%)
5–9	2710 (16%)	1937 (11%)	1316 (7%)	1329 (16%)	768 (11%)	203 (8%)
≥10	1819 (11%)	1292 (7%)	729 (4%)	804 (10%)	439 (6%)	98 (4%)
Median (IQR)	1 (0–5)	0 (0–3)	0 (0–1)	1 (0–5)	0 (0–2)	0 (0–1)
Number of opioid prescriptions						
0	8053 (49%)	9546 (53%)	11,175 (60%)	4338 (51%)	4011 (57%)	1577 (60%)
I	1962 (12%)	2042 (11%)	1951 (11%)	962 (11%)	775 (11%)	296 (11%)
2–4	2276 (14%)	2268 (13%)	1954 (11%)	1081 (13%)	824 (12%)	297 (11%)
5–9	2025 (12%)	1845 (10%)	1574 (8%)	1017 (12%)	703 (10%)	232 (9%)
≥10	2173 (13%)	2343 (13%)	1896 (10%)	1053 (12%)	772 (11%)	229 (9%)
Median (IQR)	1 (0–5)	0 (0–4)	0 (0–2)	0 (0–4)	0 (0–3)	0 (0–2)
Number of glucocorticoid prescriptions						
0	10,569 (64%)	11,241 (62%)	12,215 (66%)	5330 (63%)	4348 (61%)	1663 (63%)
I	1244 (8%)	1578 (9%)	1757 (9%)	762 (9%)	753 (11%)	297 (11%)
2–4	1425 (9%)	1801 (10%)	1975 (11%)	818 (10%)	773 (11%)	334 (13%)
5–9	1646 (10%)	1645 (9%)	1422 (8%)	942 (11%)	737 (10%)	201 (8%)
≥10	1605 (10%)	1779 (10%)	1181 (6%)	599 (7%)	474 (7%)	136 (5%)
Median (IQR)	0 (0–2)	0 (0–2)	0 (0–1)	0 (0–2)	0 (0–2)	0 (0–2)
Number of supporting clinical codes[c]						
0	9544 (58%)	4011 (22%)	1459 (8%)	3807 (45%)	1135 (16%)	286 (11%)
I	2505 (15%)	2493 (14%)	2661 (14%)	1219 (14%)	848 (12%)	358 (14%)
2–4	2677 (16%)	5486 (30%)	6365 (34%)	2166 (26%)	2584 (36%)	992 (38%)
5–9	1267 (8%)	4596 (25%)	6087 (33%)	1079 (13%)	2069 (29%)	813 (31%)
≥10	496 (3%)	1458 (8%)	1978 (11%)	180 (2%)	449 (6%)	182 (7%)
Median (IQR)	0 (0–2)	3 (1–6)	4 (2–7)	1 (0–3)	3 (1–6)	3 (2–6)

Notes: [a]RA-specific labs included are rheumatoid factor and anti-CCP antibody. General RA labs included are antinuclear antibody (ANA), C-reactive protein (CRP), human leukocyte antigen (HLA) and erythrocyte sedimentation rate. [b]Analgesics include NSAIDs and opioids. [c]Supporting clinical codes include rheumatology referrals and visits, drug monitoring, and joint pain codes.

similar, but patients in CPRD Aurum more often had 10 or more of these lab events recorded (Table 3). Patients in CPRD Aurum also more often had 10 or more prescriptions for conventional disease-modifying antirheumatic drugs (cDMARDs) after first RA diagnosis compared to patients in CPRD GOLD (Table 3).

In the year before and after first RA diagnosis, the distribution of types of recorded RA labs were similar between CPRD Aurum and CPRD GOLD, with some minor exceptions (Table 4). Patients in CPRD GOLD more often had no labs recorded in 2015–2019 in both the year before (CPRD Aurum 18%; CPRD GOLD 22%) and after first RA diagnosis (CPRD Aurum 30%; CPRD GOLD 34%) and patients in CPRD Aurum had C-reactive protein recorded more often in 2015–2019 both in the year before (CPRD Aurum 73%; CPRD GOLD 68%) and after (CPRD Aurum 63%; CPRD GOLD 59%) first RA diagnosis.

In Table 5 we present the crude and age- and sex-standardized incidence of RA (defined as presence of an RA diagnostic code) by time period in CPRD Aurum and CPRD GOLD. The overall standardized incidences were 41.56 (95%CI: 41.20–41.92) and 43.89 (95%CI: 44.89–46.22) per 100,000 person-years for CPRD Aurum and CPRD GOLD, respectively. The standardized incidence of RA was slightly higher in CPRD GOLD in the earlier time periods, but the incidence was more similar in the 2015–2019 time period – CPRD Aurum 43.48 per 100,000 person-years (95%CI: 42.86–44.12); CPRD GOLD 40.30 per 100,000 person-years (95%CI: 39.77–41.87).

Table 4 Proportion of Patients with Lab Tests Done (Not Results) Within One Year Before and on or Within One Year After First RA Diagnosis, by Data Source and Time Period (of First RA Diagnosis)

	CPRD Aurum			CPRD GOLD		
	2005–2009	2010–2014	2015–2019	2005–2009	2010–2014	2015–2019
	N=16,489 (%)	N=18,044 (%)	N=18,550 (%)	N=8,451 (%)	N=7,085 (%)	N=2,631 (%)
Before first RA diagnosis						
No RA labs	4475 (27%)	3475 (19%)	3360 (18%)	2242 (27%)	1379 (19%)	580 (22%)
Rheumatoid factor	7798 (47%)	9453 (52%)	9088 (49%)	4232 (50%)	3805 (54%)	1336 (51%)
Anti-CCP antibodies	227 (1%)	1350 (7%)	2955 (16%)	135 (2%)	651 (9%)	346 (13%)
C-reactive protein	8500 (52%)	12,263 (68%)	13,547 (73%)	4487 (53%)	4839 (68%)	1798 (68%)
Antinuclear antibody (ANA)	249 (2%)	436 (2%)	343 (2%)	165 (2%)	276 (4%)	94 (4%)
Human leucocyte antigen (HLA)	58 (0%)	95 (1%)	93 (1%)	50 (1%)	48 (1%)	24 (1%)
Erythrocyte sedimentation rate	9754 (59%)	11,684 (65%)	11,582 (62%)	5068 (60%)	4605 (65%)	1649 (63%)
On or after first RA diagnosis						
No RA labs	5462 (33%)	4563 (25%)	5506 (30%)	2996 (35%)	1994 (28%)	897 (34%)
Rheumatoid factor	2336 (14%)	1654 (9%)	1138 (6%)	1104 (13%)	628 (9%)	191 (7%)
Anti-CCP antibodies	237 (1%)	542 (3%)	447 (2%)	128 (2%)	278 (4%)	91 (3%)
C-reactive protein	8,423 (51%)	11,824 (66%)	11,766 (63%)	4200 (50%)	4444 (63%)	1553 (59%)
Antinuclear antibody (ANA)	100 (1%)	96 (1%)	43 (0%)	44 (1%)	68 (1%)	14 (1%)
Human leucocyte antigen (HLA)	45 (0%)	54 (0%)	58 (0%)	36 (0%)	32 (0%)	12 (0%)
Erythrocyte sedimentation rate	9187 (56%)	10,716 (59%)	9319 (50%)	4455 (53%)	3994 (56%)	1,274 (48%)

Note: Labs are not mutually exclusive.

Table 5 Crude and Standardized Incidence Rates of RA (Defined as Presence of at Least 1 RA Diagnostic Code) per 100,000 Person-years, by Data Source and Time Period (of First RA Diagnosis), in England

	CPRD Aurum			CPRD GOLD		
	2005–2009	**2010–2014**	**2015–2019**	**2005–2009**	**2010–2014**	**2015–2019**
Total RA cases	16,489	18,044	18,550	8451	7085	2631
Denominator person-years	44,084,277	47,130,368	45,455,856	18,367,802	16,283,251	6,737,602
Overall crude incidence rate (95%CI)	37.40 (36.84–38.98)	38.29 (37.73–38.85)	40.81 (40.22–41.40)	46.01 (45.04–47.00)	43.51 (42.51–44.53)	39.05 (37.58–40.56)
Overall age- and sex- standardized[a] incidence rate (95%CI)	40.24 (39.62–40.86)	40.98 (40.39–41.58)	43.48 (42.86–44.12)	47.98 (46.95–49.02)	45.04 (43.99–46.10)	40.30 (39.77–41.87)

Note: [a]Standardized using Office for National Statistics 2018 age- and sex-specific population estimates.[25]

Discussion

In our previous short report, we found that codes expected for monitoring and treatment of RA were routinely recorded in CPRD Aurum.[21] Our current study provides a more thorough description of RA-related data captured in the CPRD Aurum patient record, as well as a side-by-side comparison to the equivalent data recorded in CPRD GOLD. Overall, the recording of diagnoses, prescription drugs and labs in the year before and the year after first RA diagnosis were similar between CPRD Aurum and CPRD GOLD. Slight differences were found for a few of the variables under study in some time periods, but overall, we found consistency between the two databases, particularly in more recent years. In addition, the incidence of RA, standardized to the Office for National Statistic's 2018 data, was similar between CPRD Aurum and CPRD GOLD. These rates were somewhat higher than those reported in other studies of incidence of RA due to the inclusion of all patients with an RA diagnosis, without requiring supporting codes or drug treatment.[23,26–28] However, the objective of this study was an assessment of CPRD Aurum data quality and completeness, not complete capture of RA cases; rates presented in this study should not be considered unbiased estimates of RA incidence in CPRD Aurum or CPRD GOLD.

This study has several limitations. CPRD Aurum and CPRD GOLD use different diagnosis and drug codes to record data (Supplemental Table 1). Though we tried to match the diagnosis, drug, and lab codes between the two databases using text string searches of the available codes, it is possible that code lists were more complete for one database than the other. Given the consistency between the two databases in recording of diagnoses, drugs, and labs, as well as similar incidence rates, it is unlikely that we missed important RA codes. We could not assess differences in use of biologics between CPRD Aurum and CPRD GOLD because there is limited information on these treatments in both databases as these drugs are not typically prescribed by the GP.[29] We looked at the presence of lab tests, but we did not evaluate the lab results. We did not assess differences in the recording of medical care over the progression of disease because we limited our analyses to the time around the first RA diagnosis.

CPRD GOLD was created as an electronic health record with the understanding that the data would be used in medical research. As such, the GPs who originally signed on to contribute to CPRD GOLD were provided training in data recording, whereas contributors to CPRD Aurum were not provided this training. It is possible that differences seen between recording of data in these two databases may be a result of the difference in data recording training, but as this training occurred over 30 years ago, and electronic health records are now the norm, it is unlikely to have resulted in meaningful differences in data quality in the years covered in this study.

Conclusion

Over 30 years, CPRD GOLD has been used for thousands of studies. As patient numbers in CPRD GOLD decline, CPRD Aurum data has become available for research within a similar population, and with ever increasing numbers of patients. If CPRD Aurum is used by researchers as a replacement or in addition to CPRD GOLD, it is important that researchers understand the data and how it compares to CPRD GOLD. This study shows that among patients with RA, relevant variables used to diagnose and monitor RA are consistently recorded between CPRD Aurum and CPRD GOLD.

The information in CPRD Aurum is generally comparable to information recorded in CPRD GOLD, including recording of diagnoses, drugs, and labs. In particular, incidence of RA was similar between the two databases, suggesting that the recording of RA in CPRD Aurum does not miss a large number of RA cases. As more researchers use CPRD Aurum, additional validation studies should be conducted for other chronic diseases.

Acknowledgments

Funding

Disclosure

Dr Catherine Vasilakis-Scaramozza, Ms Katrina Wilcox Hagberg, Ms Rebecca Persson, and Dr Susan Jick report that the BCDSP (employer) received contractual research grants from Amgen during the conduct of the study. Dr George Kafatos, Dr Joe Maskell, and Dr David Neasham 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 report no other conflicts of interest in this work.

References

1. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol*. 2019;48 (6):1740–1740g. doi:10.1093/ije/dyz034
2. 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(9):1–7.
3. 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;13:745–753. doi:10.2147/CLEP.S319245
4. 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
5. 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 Epidemiol*. 2022;6:6. doi:10.21037/ace-22-4
6. Cybulski L, Ashcroft DM, Carr MJ, et al. Temporal trends in annual incidence rates for psychiatric disorders and self-harm among children and adolescents in the UK, 2003-2018. *BMC Psychiatry*. 2021;21(1):229. doi:10.1186/s12888-021-03235-w
7. Davidson J, Banerjee A, Mathur R, et al. Ethnic differences in the incidence of clinically diagnosed influenza: an England population-based cohort study 2008-2018. *Wellcome Open Res*. 2021;6:49. doi:10.12688/wellcomeopenres.16620.3
8. Hayes JF, Osborn DPJ, Francis E, et al. Prediction of individuals at high risk of chronic kidney disease during treatment with lithium for bipolar disorder. *BMC Med*. 2021;19(1):99. doi:10.1186/s12916-021-01964-z
9. Lewer D, Padmanathan P, Qummer Ul Arfeen M, et al. Healthcare use by people who use illicit opioids (HUPIO): development of a cohort based on electronic primary care records in England. *Wellcome Open Res*. 2021;5:282. doi:10.12688/wellcomeopenres.16431.2
10. Gulliford MC, Sun X, Anjuman T, et al. Comparison of antibiotic prescribing records in two UK primary care electronic health record systems: cohort study using CPRD GOLD and CPRD Aurum databases. *BMJ Open*. 2020;10(6):e038767.
11. Requena G, Wolf A, Williams R, et al. Feasibility of using Clinical Practice Research Datalink data to identify patients with chronic obstructive pulmonary disease to enrol into real-world trials. *Pharmacoepidemiol Drug Saf*. 2021;30(4):472–481. doi:10.1002/pds.5188
12. 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.
13. Fischer LM, Schlienger RG, Matter C, et al. Effect of rheumatoid arthritis or systemic lupus erythematosus on the risk of first-time myocardial infarction. *Am J Cardiol*. 2004;93:198–200. doi:10.1016/j.amjcard.2003.09.037
14. Hawley S, Cordtz R, Dreyer L, et al. Association between NICE guidance on biologic therapies with rates of Hip and knee replacement among rheumatoid arthritis patients in England and Wales: an interrupted time-series analysis. *Semin Arthritis Rheum*. 2018;47(5):605–610. doi:10.1016/j.semarthrit.2017.09.006
15. Jick SS, Choi H, Li L, et al. Hyperlipidemia, statin use and the risk of developing rheumatoid arthritis. *Ann Rheum Dis*. 2009;68:546–551. doi:10.1136/ard.2008.091967
16. Klop C, de Vries F, Bijlsma JW, et al. Predicting the 10-year risk of Hip and major osteoporotic fracture in rheumatoid arthritis and in the general population: an independent validation and update of UK FRAX without bone mineral density. *Ann Rheum Dis*. 2016;75(12):2095–2100. doi:10.1136/annrheumdis-2015-208958
17. Movahedi M, Beauchamp ME, Abrahamowicz M, et al. Risk of incident diabetes mellitus associated with the dosage and duration of oral glucocorticoid therapy in patients with rheumatoid arthritis. *Arthritis Rheumatol*. 2016;68(5):1089–1098. doi:10.1002/art.39537

18. Muller S, Hider SL, Raza K, et al. An algorithm to identify rheumatoid arthritis in primary care: a Clinical Practice Research Datalink study. *BMJ Open*. 2015;5(12):e009309. doi:10.1136/bmjopen-2015-009309

19. Nagra NS, Robinson DE, Douglas I, et al. Antibiotic treatment and flares of rheumatoid arthritis: a self-controlled case series study analysis using CPRD GOLD. *Sci Rep*. 2019;9(1):8941. doi:10.1038/s41598-019-45435-1

20. Sultan AA, Mallen C, Muller S, et al. Antibiotic use and the risk of rheumatoid arthritis: a population-based case-control study. *BMC Med*. 2019;17 (1):154. doi:10.1186/s12916-019-1394-6

21. 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

22. 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

23. National Institute for Health and Care Excellence. Rheumatoid arthritis in adults: management (Clinical guideline [CG79]); 2009. Available from: https://www.nice.org.uk/guidance/cg79. Accessed December 06, 2023.

24. National Institute for Health and Care Excellence. Rheumatoid arthritis in adults: management (Clinical guideline [NG100]); 2018. Available from: https://www.nice.org.uk/guidance/ng100. Accessed December 06, 2023.

25. 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 December 06, 2023.

26. Humphreys JH, Verstappen SM, Hyrich KL, et al. The incidence of rheumatoid arthritis in the UK: comparisons using the 2010 ACR/EULAR classification criteria and the 1987 ACR classification criteria. Results from the Norfolk Arthritis Register. *Ann Rheum Dis*. 2013;72(8):1315–1320. doi:10.1136/annrheumdis-2012-201960

27. Russell MD, Galloway JB, Andrews CD, et al. OpenSAFELY Collaborative. Incidence and management of inflammatory arthritis in England before and during the COVID-19 pandemic: a population-level cohort study using OpenSAFELY. *Lancet Rheumatol*. 2022;4(12):e853–e863. doi:10.1016/S2665-9913(22)00305-8

28. Scott IC, Whittle R, Bailey J, et al. Rheumatoid arthritis, psoriatic arthritis, and axial spondyloarthritis epidemiology in England from 2004 to 2020: an observational study using primary care electronic health record data. *Lancet Reg Health Eur*. 2022;23:100519. doi:10.1016/j.lanepe.2022.100519

29. Okoli GN, Myles P, Murray-Thomas T, et al. Use of primary care data in research and pharmacovigilance: eight scenarios where prescription data are absent. *Drug Saf*. 2021;22:1–8.