Clinical Epidemiology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Comparison of Epidemiological Estimates for 19 Cancer Types Using Electronic Health Record Databases in England: An Analysis of CPRD Aurum and CPRD GOLD Databases with Linked Hospital Episode Statistics and Cancer Registry Data

Anna B Chaplin[1], Olia Archangelidi[1], Katrina Wilcox Hagberg[2], David Neasham[1], George Kafatos[1]

[1]Center for Observational Research, Amgen Limited, Uxbridge, UK; [2]Boston Collaborative Drug Surveillance Program, Lexington, MA, USA

Correspondence: Anna B Chaplin, Center for Observational Research, Amgen Ltd, Uxbridge Business Park, 4 Sanderson Road, Denham, Uxbridge, UB8 1DH, UK, Tel +44 0 7741 601 593, Email achapl01@amgen.com

Purpose: To compare epidemiological estimates of 19 different cancer types in Clinical Practice Research Datalink (CPRD) Aurum and CPRD General Practice Online Database (GOLD) databases against linked secondary data sources in England to understand best use of these data sources for research.

Methods: The source population comprised patients in CPRD Aurum or GOLD (separately) who were eligible for linkage to Cancer Registry (CR), Hospital Episode Statistics (HES), and Office for National Statistics (ONS). We selected patients who had an incident cancer diagnosis recorded in ≥1 data sources (CPRD Aurum or GOLD, HES, CR) between January 1, 2011 and December 31, 2018. We estimated incidence rates (IR) and counts by cancer type and data source, and survival probability by cancer type among patients with an ONS death record recorded between January 1, 2011 and April 30, 2020.

Results: The highest incident case capture resulted from CPRD Aurum or GOLD linked to HES and CR. In the fully linked CPRD Aurum-HES-CR and CPRD GOLD-HES-CR datasets, cancers typically diagnosed and managed in primary care (eg, breast, prostate, and lung) had the highest IRs and more complete case capture compared with other data sources, whereas HES and CR had higher IRs for cancers diagnosed in secondary care settings (eg, gastric, renal, and bladder). Cancers with broad definitions (eg, head and neck) had wider variations in IRs across data sources than cancers with narrower definitions. Survival estimates were generally higher for cancer-related deaths versus all-cause deaths.

Conclusion: Findings highlight variation in cancer recording across different data sources. Researchers using CPRD data should assess the benefit of incorporating linked data on a study-by-study basis. For studies of breast, prostate, and lung cancers, CPRD Aurum or GOLD alone may be sufficient; however, linkage to HES and/or CR is recommended where a more complete case capture is required.

Plain Language Summary:

Purpose: To compare the number of patients with 19 different cancer types recorded in databases from general practitioner (GP) practices in England with other databases. The aim was to help researchers understand and interpret results from these databases.

Methods: The GP practices databases, Clinical Practice Research Datalink (CPRD) Aurum and GOLD, were compared with the Cancer Registry (CR), Hospital Episode Statistics (HES), and Office for National Statistics (ONS) death records. Patients in CPRD Aurum or GOLD (separately) were included in the study if their medical record could be linked to CR, HES, and ONS. We selected patients who had their first ever (incident) cancer diagnosis recorded in ≥1 data sources between January 1, 2011 and December 31, 2018. We estimated the number and rate of new cancer diagnoses by cancer type and data source. We estimated the probability of survival by cancer type among patients with a death record between January 1, 2011 and April 30, 2020.

Results: The number of new cancer cases varied depending on the type of cancer. Linking GP records to hospital and cancer registry data resulted in the highest count of new cancer cases. GP data alone captured more diagnoses for cancers typically diagnosed and managed in primary care, while CR and HES linked datasets had higher rates for cancers diagnosed in secondary care.

Received: 5 August 2025
Accepted: 14 November 2025
Published: 6 December 2025

Conclusion: The results show that cancer records differ between various data sources. Researchers using CPRD data should carefully consider whether to use linked data for each specific study.

Keywords: CPRD Aurum, CPRD GOLD, UK cancer registries, Hospital Episode Statistics, cancer incidence

Introduction

The Clinical Practice Research Datalink (CPRD) offers access to two medical databases, CPRD Aurum and CPRD GOLD, containing deidentified primary care electronic health records of patients from a network of general practitioners (GPs) in the United Kingdom (UK).[1,2] While both databases are widely used for research, differences in their coverage, coding systems, and time periods may affect the completeness and accuracy of cancer diagnosis recording.[3,4] CPRD Aurum covers English primary care practices while GOLD covers all UK primary care practices.[5,6] The health information available in CPRD Aurum and GOLD can be enriched through linkage to the National Cancer Registration and Analysis Service Cancer Registry (CR) data, Hospital Episode Statistics (HES) Admitted Patient Care data, and the Office for National Statistics (ONS) death registration data.[7] These linkages provide the opportunity to assess the scale of missing cancer diagnoses through the comparison of cancer diagnoses recorded in primary care data (CPRD Aurum and GOLD) with those recorded in CR and HES data.

Several studies have reported that clinical information captured in CPRD Aurum and GOLD is of high accuracy and completeness regarding several clinical conditions, pharmacological prescriptions, and deaths.[4,8–14] However, assessments of cancer diagnosis records have yielded different results, with the capture of cancer cases varying by cancer type.[3,12,15–19] In a study of 116,769 patients, records from CPRD GOLD with linked HES were compared with CR data for five cancer types.[16] Around 10% of cases identified from CPRD GOLD or HES did not have a confirmatory diagnosis recorded in CR, while up to 32% of other cancer cases identified in linked CR were missing in CPRD GOLD.[16] Linking HES and CR to CPRD Aurum or GOLD for studies on certain cancer types may be warranted to improve the accuracy and completeness of case capture, and should be considered for studies requiring data that are not captured in GP records, such as cancer stage or grade.[20]

Previous studies comparing cancer diagnosis information across medical databases have typically focused on a few cancer types. Our study aimed to comprehensively describe how well cancer diagnoses are recorded across CPRD Aurum and GOLD for 19 cancer types, and to evaluate cancer recording in linked data to describe best use of these data resources for cancer research. The objectives were to: (i) measure and compare the incidence of 19 cancer types calculated based on CPRD Aurum and GOLD databases with and without linkage to CR and HES; and (ii) estimate survival by cancer type using ONS death registry data.

By comparing the completeness and accuracy of cancer diagnosis recording across multiple data sources and cancer types, this study aims to offer new insights into the strengths and limitations of primary care data for cancer research. The findings could help to guide researchers in selecting the most appropriate data sources for future studies, and to highlight the value of data linkage in improving case capture and epidemiological estimates.

Methods

Study Design

This retrospective cohort study estimated incidence rates (IRs) and survival probabilities of 19 cancer types annually among patients in England. Estimates were described separately for CPRD Aurum and GOLD, linked with HES, CR, and ONS death registry to cross-reference cancer diagnoses across the different data sources.

Data Sources

CPRD Aurum and GOLD contain deidentified primary care data, including demographics and diagnoses.[5,6] CPRD Aurum, launched in 2017, covers 24.75% of the UK population. It is based on primary care data contributed from practices in England collected using the EMIS Web® software, with data available from 1995 onwards.[5,21] CPRD GOLD, formerly known as the General Practice Research Database, was established in 1987 and covers 4.17% of the UK population, with data available

from 1987 onwards.[22] It is based on primary care data captured from practices in Scotland, Wales, Northern Ireland, and England via the Vision™ patient management software.[5,6]

CPRD data for patients in England can be linked to HES, CR, and ONS data.[7] HES provides hospital episode data in England from 1997 onwards, including admission and discharge dates, and diagnoses.[7] CR data contain every registrable tumor diagnosed or treated in England.[7,23] The ONS contains data pertaining to the official date and cause of death for people in England. In relation to the latter, it is a legal requirement to certify and register all deaths in England, so these data can be regarded as well ascertained.[7,24]

Study Population

Cancer types included were acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), bladder cancer, brain cancer, breast cancer, colorectal cancer, esophageal cancer, gastric cancer, head and neck cancer, lung cancer, melanoma, multiple myeloma (MM), neuroendocrine cancer, ovarian cancer, pancreatic cancer, prostate cancer, renal cancer, thyroid cancer, and uterine cancer. Cancer diagnoses and death were identified using International Classification of Diseases 10th Revision (ICD-10) codes, CPRD Aurum medical codes, and CPRD GOLD diagnostic codes. Diagnostic codes for included cancer types can be found in Supplementary Table 1. The analysis of prostate cancer was restricted to male patients, while that of ovarian and uterine cancers was restricted to female patients.

Patient selection criteria were applied separately for CPRD Aurum and GOLD. Patients were included if they: (i) fulfilled the CPRD data-quality standards (ie, "acceptable" flag) for CPRD Aurum and GOLD; (ii) were eligible for linkage to HES, CR, or ONS data; and (iii) were registered in the CPRD database for at least one day during the study period (January 1, 2011 to December 31, 2018). Patients eligible for linkage were those from GP practices in England who consented to participate in the linkage process, who had not opted out from sharing their records, and whose primary care records had a valid identifier to allow for linkage to HES, CR, or ONS death registration data.[7] Since this study was restricted to patients in CPRD Aurum or GOLD who were eligible for linkage to HES, CR, or ONS, the study population was restricted to patients in England only, based on linkage availability.

Patients with an incident cancer diagnosis in at least one of the data sources (CPRD, HES, CR) between January 1, 2011 (or the GP registration start date, whichever was latest) and December 31, 2018 (or GP registration end date or last collection date of the GP practice, whichever was earliest) were included for IR analysis (Figure 1). An incident diagnosis was defined as the first recorded diagnosis code of a particular cancer type (a patient may have had more than one incident cancer type diagnosed). This approach may result in misclassification, whereby recurrent cancers could be incorrectly classified as incident cases, particularly in the absence of complete historical data. This methodology, however, is consistent with previous large-scale epidemiological studies using UK electronic health records, where similar definitions have been adopted due to data structure limitations.[16,25,26] Patients with an ONS death record during the follow-up period (January 1, 2011 to April 30, 2020) were included for survival analysis.

To evaluate the completeness of cancer diagnoses (as estimated by IRs) recorded in the various combinations of data sources, we created separate analytical datasets for patients whose cancer diagnoses were identified in CPRD Aurum or GOLD and their data linkages: (i) CPRD Aurum patients: CPRD Aurum records only; (ii) CPRD Aurum patients: HES records only; (iii) CPRD Aurum patients: CR records only; (iv) CPRD Aurum patients: HES-CR records; (v) CPRD Aurum patients: CPRD Aurum-HES-CR records (fully linked data); (vi) CPRD GOLD patients: CPRD GOLD records only; (vii) CPRD GOLD patients: HES records only; (viii) CPRD GOLD patients: CR records only; (ix) CPRD GOLD patients: HES-CR records; and (x) CPRD GOLD patients: CPRD GOLD-HES-CR records (fully linked data). These datasets are further described in Table 1.

To compare the completeness of cancer diagnosis records, we compared cancer types coded in CPRD Aurum and GOLD with those in the respective fully linked datasets (CPRD Aurum-HES-CR or CPRD GOLD-HES-CR). The CPRD, HES, or CR cancer diagnosis code appearing first on a patient's record was taken as the incident diagnosis for that cancer type in the fully linked dataset for patients in the CPRD Aurum and GOLD databases separately. The fully linked datasets were used as references for this analysis, as it was assumed that combining the GP databases with linked HES and CR data would result in the most complete cancer-case capture.

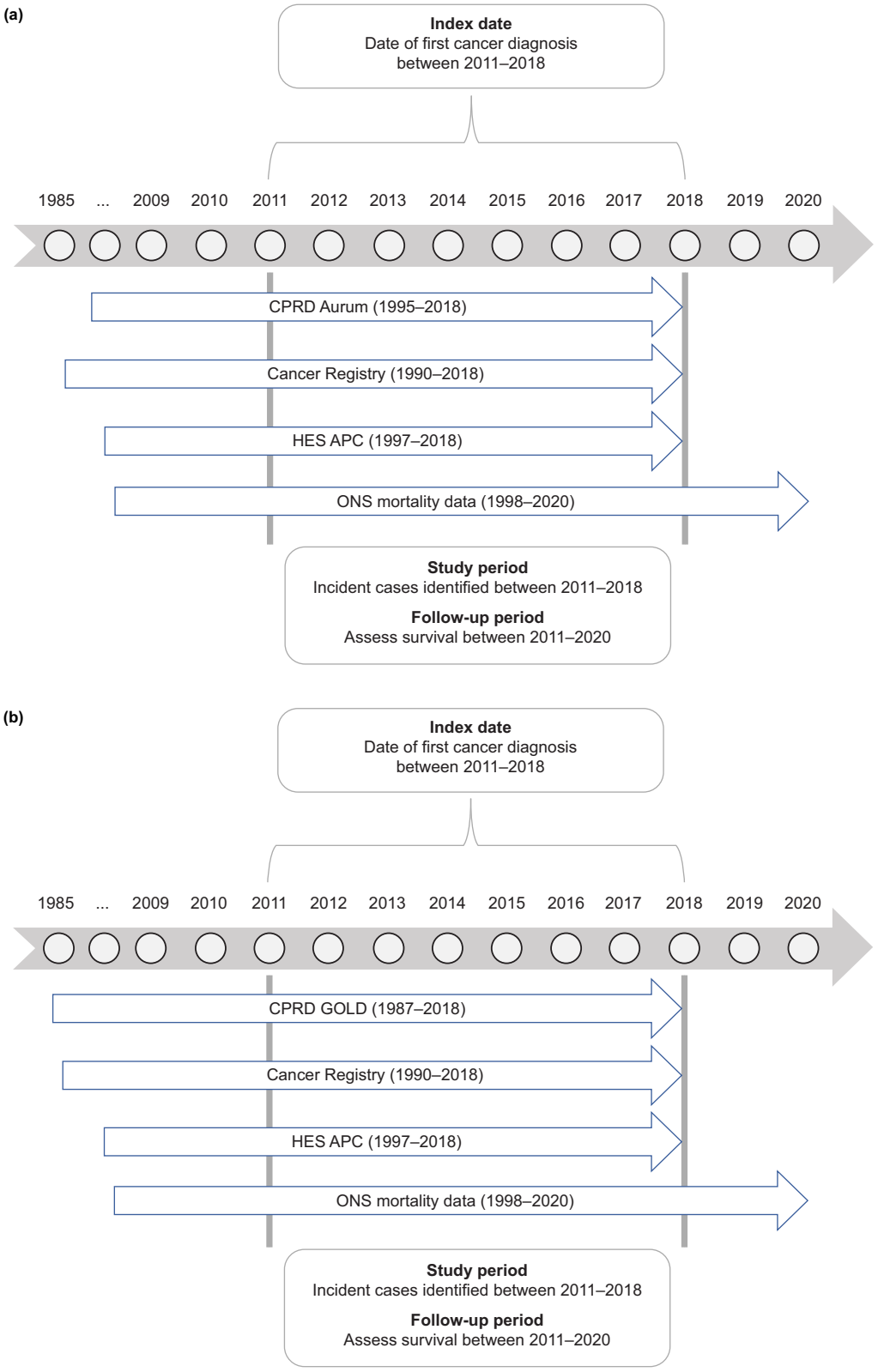

Figure 1 Study design and data sources (**a**) CPRD Aurum and (**b**) CPRD GOLD.
Abbreviations: APC, Admitted Patient Care; CPRD, Clinical Practice Research Datalink; HES, Hospital Episode Statistics; ONS, Office for National Statistics.

Table 1 Analytical Dataset Selection for Incidence Rate Analysis and Cross-Comparison of Diagnoses by Data Source

Dataset Name	Cancer Cases (Numerator)	Source Population (Denominator)
CPRD Aurum study population		
CPRD Aurum records only	All CPRD Aurum source population patients with an incident cancer diagnosis recorded in CPRD Aurum	All patients in CPRD Aurum eligible for linkage to HES, CR, or ONS (England only)
CPRD Aurum-HES-CR (fully linked, reference)	All CPRD Aurum source population patients with an incident cancer diagnosis recorded in CPRD Aurum, HES, and/or CR	
CR records only	All CPRD Aurum source population patients with an incident cancer diagnosis recorded in CR	
HES records only	All CPRD Aurum source population patients with an incident cancer diagnosis recorded in HES	
HES-CR records	All CPRD Aurum source population patients with an incident cancer diagnosis recorded in CR and/or HES	
CPRD GOLD study population		
CPRD GOLD records only	All CPRD GOLD source population patients with an incident cancer diagnosis recorded in CPRD Aurum	All patients in CPRD GOLD eligible for linkage to HES, CR, or ONS (England only)
CPRD GOLD-HES-CR (fully linked, reference)	All CPRD GOLD source population patients with an incident cancer diagnosis recorded in CPRD Aurum, HES, and/or CR	
CR records only	All CPRD GOLD source population patients with an incident cancer diagnosis recorded in CR	
HES records only	All CPRD GOLD source population patients with an incident cancer diagnosis recorded in HES	
HES-CR records	All CPRD GOLD source population patients with an incident cancer diagnosis recorded in CR and/or HES	

Abbreviations: CPRD, Clinical Practice Research Datalink; CR, Cancer Registry; HES, Hospital Episode Statistics; ONS, Office for National Statistics.

The fully linked datasets were also used for survival analysis. Patients were included in the survival analysis if they had at least one diagnosis of one of the 19 prespecified cancer types recorded in at least one of the study datasets. Death records were taken from the ONS database. A follow-up period from January 1, 2011 to April 30, 2020 was used to capture survival. Many factors that influence survival, such as cancer stage, histology, and treatment regimen, were not evaluated during this study. As such, survival analyses are unadjusted.

Statistical Analysis

Analyses conducted included incidence counts, IRs, proportion comparisons, and survival probability estimates, all by cancer type. Estimates were calculated separately for CPRD Aurum and GOLD linked datasets.

IRs per cancer type were estimated annually (2011 to 2018) for each of the 10 analytical datasets. The IR (per 100,000 person-years) of cancer cases per year per cancer type was estimated as the number of new cancer cases divided by the total number of person-years (see Table 1). New cases were defined as patients with their first diagnosis of that type of cancer. A patient was classed as an incident case in a particular year if their first diagnosis occurred in their patient record within that particular year and they had a record in the relevant database for one or more days within that year. The denominator included the total person-years of patients registered in CPRD Aurum or GOLD (separately) per year of the study period. For each IR estimate, 95% confidence intervals were computed using the Poisson distribution. Comparisons between IRs across datasets and cancer types were performed descriptively. All statistical analyses were conducted using SAS version 9.4 (SAS Institute, Cary, NC).

The proportion of cancer cases captured by each individual dataset (CPRD Aurum/GOLD, HES, CR) in comparison with the cases recorded in the fully linked dataset (CPRD Aurum- or GOLD-HES-CR) was calculated to assess the capture of incident outcomes in each data source.

Kaplan–Meier survival probabilities over time were estimated among patients diagnosed with cancer between January 1, 2011 and December 31, 2018 by cancer type. Patients were followed up from incident diagnosis until death, loss to follow-up, or end of the follow-up period (April 30, 2020). Death from any cause and cancer-related death (ie, where cancer was listed as the primary cause of death such that the main ICD-10 code in ONS data corresponded to that specific cancer type) were outcomes of interest. Outcomes were presented as separate lines on the same Kaplan–Meier plot for each cancer type. The fully linked dataset (CPRD-HES-CR) was linked to ONS death data to calculate unadjusted survival probabilities.

Owing to the structure of the available data, patient time at risk was not censored at the time of event occurrence, and person-time denominators included all time registered in the database. This methodological limitation may result in underestimation of IRs and should be considered when interpreting the results.

Results

Details of the patient sample size and attrition for each population dataset are set out in Table 2. The incident cancer counts for the fully linked CPRD Aurum-HES-CR and CPRD GOLD-HES-CR datasets are shown in Supplementary Table 2. The case counts in the CPRD GOLD-HES-CR dataset declined over time and increased in the CPRD Aurum-HES-CR dataset due to the gradual replacement of the Vision patient management software by the EMIS system.[5]

Table 2 Patient Numbers and Attrition in Population Datasets

CPRD Aurum	**n**
CPRD Aurum alone	
Patients in the CPRD Aurum dataset	1,123,375
Patients with an observation record	1,123,352
Patients with a cancer record before 2019	649,504
Patients registered during the study period	649,504
HES alone	
Patients with a diagnosis in the HES Aurum database	1,068,261
Patients with a primary diagnosis of cancer before 2019	632,309
Patients registered during the study period	632,309
CR alone	
Patients with a record in the CR Aurum database	633,507
Patients with a cancer record before 2019	552,623
Patients registered during the study period	552,623
HES-CR	
Patients who met criteria for either HES or CR cohorts	747,961
Aurum-HES-CR	
Patients who met criteria for CPRD, HES, or CR cohorts	845,939

(*Continued*)

Table 2 (Continued).

CPRD GOLD	n
CPRD GOLD alone	
Patients in the CPRD GOLD dataset	305,512
Patients with a diagnosis record	305,408
Patients with a cancer record before 2019	153,844
Patients registered during the study period	147,473
HES alone	
Patients with a diagnosis in the HES GOLD database	298,357
Patients with a primary diagnosis of cancer before 2019	192,218
Patients registered during the study period	185,208
CR alone	
Patients with a record in the CR GOLD database	267,666
Patients with a cancer record before 2019	224,164
Patients registered during the study period	216,202
HES-CR	
Patients who met criteria for either HES or CR cohorts	231,899
GOLD-HES-CR	
Patients who met criteria for CPRD, HES, or CR cohorts	248,587

Abbreviations: CPRD, Clinical Practice Research Datalink; CR, Cancer Registry; HES, Hospital Episode Statistics.

Cancer Type-Specific Incidence Rates

CPRD Aurum and Linked Datasets

Overall, the cancer type-specific IRs in CPRD Aurum and the linked datasets were generally similar regardless of data source (Figure 2a). IRs were similar on an absolute scale; however, differences on a relative scale were more pronounced for rarer cancers. The highest IRs for all cancer types, suggesting a more complete case capture (assuming no false positives), were obtained using the fully linked CPRD Aurum-HES-CR dataset. IRs for prostate, breast, lung, and colorectal cancer were the highest during the study period. The IR for head and neck cancer varied widely by data source used, with the lowest capture in the CPRD Aurum alone, and the highest capture in the fully linked CPRD Aurum-HES-CR datasets (Figure 2). Differences in IRs by data source appeared to be smaller for frequently reported cancer types (eg, breast cancer) or certain more aggressive cancers (eg, lung cancer). After 2013, the IR for bladder cancer based on CR data was similar to that observed in other data sources. The CPRD Aurum data alone had a more complete capture of cancers typically diagnosed and managed in primary care settings (eg, breast and prostate cancers) when compared with the linked data sources. In contrast, the IRs for cancers typically diagnosed in secondary care settings (eg, gastric, renal, bladder cancers) were higher in HES and CR linked datasets as compared with CPRD Aurum alone.

CPRD GOLD and Linked Datasets

Overall, IRs for the different cancer types in the CPRD GOLD datasets were similar to those in the CPRD Aurum datasets. The fully linked dataset (CPRD GOLD-HES-CR) had a higher case capture, as evidenced by the higher IRs regardless of cancer type (Figure 2b). The trends of IRs for the different cancer types over time in CR were similar in CPRD Aurum and GOLD.

https://doi.org/10.2147/CLEP.S558429

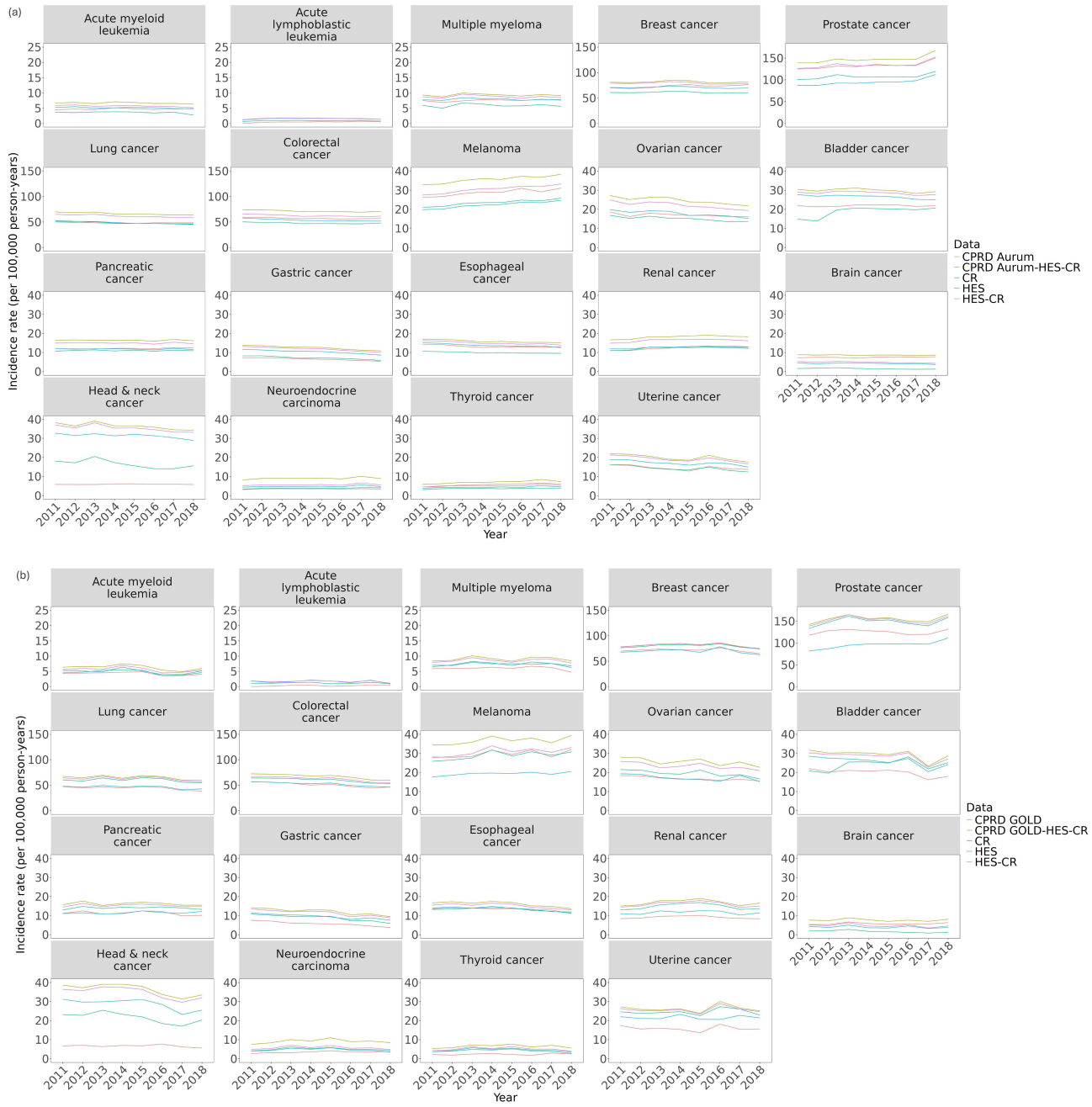

Figure 2 Incidence rates in (**a**) CPRD Aurum and linked HES and CR datasets, and (**b**) CPRD GOLD and linked HES and CR datasets over time.
Abbreviations: CPRD, Clinical Practice Research Datalink; CR, Cancer Registry; HES, Hospital Episode Statistics.

Comparison of Proportion of Incident Diagnosis by Dataset and Year

In comparison with the fully linked CPRD Aurum-HES-CR dataset, the proportion of diagnoses recorded in each dataset (CPRD Aurum, HES, CR) was relatively similar across many cancer types (Table 3 and Supplementary Figure 1a). CPRD Aurum captured a higher proportion of breast, prostate, melanoma, colorectal, pancreatic, renal, brain, and thyroid cancers compared with HES and CR datasets. HES data captured a higher proportion of AML, ALL, MM, bladder, gastric, head and neck, and uterine cancers compared with the other data sources. CR data generally had a lower proportion of cancer diagnoses recorded when compared with CPRD Aurum and/or HES, and this trend remained stable over time.

Table 3 Median Proportion of Incident Diagnoses Recorded Over Time by Data Source in the Fully Linked Analytic Datasets

	Median[a] Proportion (%) for CPRD Aurum-HES-CR			Median[a] Proportion (%) for CPRD GOLD-HES-CR		
	CPRD Aurum	HES	CR	CPRD GOLD	HES	CR
Acute lymphoblastic leukemia	29.3	95.7	55.3	21.4	96.2	70.0
Acute myeloid leukemia	70.2	77.0	54.4	68.3	79.2	75.2
Bladder cancer	72.9	89.2	67.0	69.0	87.8	84.8
Brain cancer	86.9	49.9	17.4	75.4	51.3	22.5
Breast cancer	89.6	86.8	74.9	88.1	85.6	97.9
Colorectal cancer	80.3	75.2	66.9	76.1	77.9	89.5
Esophageal cancer	83.4	86.8	62.9	81.8	84.0	82.8
Gastric cancer	52.1	83.4	57.9	47.1	79.5	73.6
Head and neck cancer	16.7	86.0	46.2	17.8	78.6	59.6
Lung cancer	73.5	72.9	72.2	68.2	71.4	92.8
Melanoma	80.5	62.5	65.8	81.0	53.0	78.1
Multiple myeloma	81.7	84.6	63.9	69.2	80.0	82.0
Neuroendocrine cancer	39.0	52.8	42.9	38.6	52.0	54.4
Ovarian cancer	69.5	72.6	61.3	66.0	68.1	76.8
Pancreatic cancer	73.9	71.9	67.2	70.5	70.9	86.8
Prostate cancer	90.6	64.1	72.8	80.1	62.4	95.9
Renal cancer	68.3	69.8	67.2	54.6	68.4	87.5
Thyroid cancer	81.5	62.3	53.7	37.1	67.3	73.1
Uterine cancer	73.4	85.8	71.7	60.4	83.6	93.2

Notes: [a]Median over time where the denominators are the number of diagnoses captured in any of the three databases (CPRD, CR, HES).
Abbreviations: CPRD, Clinical Practice Research Datalink; CR, Cancer Registry; HES, Hospital Episode Statistics.

For CPRD GOLD, the results were broadly similar to those for CPRD Aurum; however, CR captured a higher proportion of incident diagnoses for prostate, renal, thyroid, pancreatic, lung, colorectal, breast, renal, uterine, and neuroendocrine cancers (Table 3 and Supplementary Figure 1b).

Survival by Dataset and Year

For the fully linked CPRD Aurum-HES-CR dataset, any cause of death resulted in a smaller survival probability over time for all cancer types compared with cancer-related death (Figure 3a). The difference between cancer-related death and any cause of death appeared to be less prominent for more aggressive cancers (eg, pancreatic cancer). Survival probabilities for the fully linked CPRD GOLD-HES-CR dataset were consistent with those for CPRD Aurum (Figure 3b).

https://doi.org/10.2147/CLEP.S558429

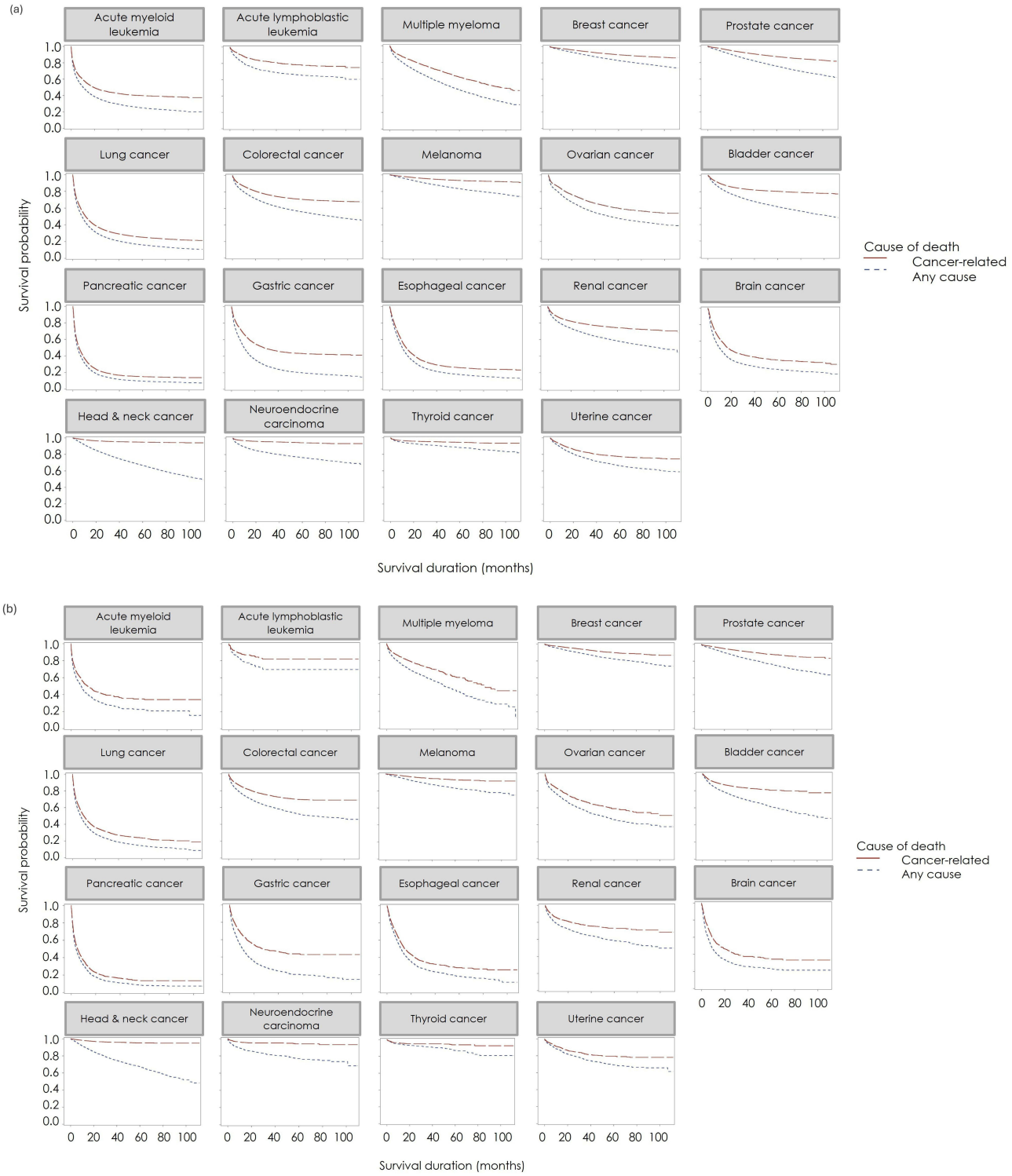

Figure 3 Kaplan–Meier plots of cancer survival probability using fully linked (**a**) CPRD Aurum data (CPRD Aurum-HES-CR), and (**b**) CPRD GOLD data (CPRD GOLD-HES-CR).
Abbreviations: CPRD, Clinical Practice Research Datalink; CR, Cancer Registry; HES, Hospital Episode Statistics.

Discussion

This study represents the most comprehensive comparison of cancer diagnosis recording, covering 19 distinct cancer types in CPRD Aurum and GOLD linked to HES and CR datasets, and provides a valuable reference for researchers navigating data

source selection for CPRD-based oncology studies. The findings demonstrate homogeneity in the recording of epidemiological estimates across CPRD Aurum and GOLD, and linked HES and CR data. Our results indicate that the highest IRs were obtained using the fully linked datasets (eg, most complete case capture). Several cancers had a lower proportion of cases coded in CPRD Aurum and GOLD, likely reflecting patient care pathways for these cancer types. While linkages are valuable, researchers should be aware of the practical impacts of using linked data, such as reduced sample size, geographic generalizability, and lag in data availability for linked data. Decisions regarding use of linkages should be based on the requirements of the study question and be balanced against practical considerations.[20]

Overall, the IRs were similar for CPRD Aurum and CPRD GOLD, and broadly in line with published research. For example, Cancer Research UK age-standardized IRs (per 100,000 person-years) between 2011 and 2018 were 5–6 for AML (vs crude IR of 3–7 in this study),[25] 10–13 for gastric cancer (vs 4–14 in this study),[26] and 23–24 for ovarian cancer (vs 13–27 in this study).[27] Variations in the epidemiological estimates of certain cancer types across the different data sources may be due to various factors, such as diagnosis at primary versus secondary care settings, prognosis, and different coding systems. As with previous studies,[4,28,29] we observed that the proportion of diagnoses captured in the different data sources varied by cancer type, which is likely a reflection of diagnostic and care pathways. For example, IRs estimated based on CPRD Aurum and CPRD GOLD alone had a high capture of diagnoses for cancer types commonly diagnosed and managed in primary care settings (eg, breast and prostate cancer), possibly because these cancer types can be treated with hormone-only treatments at first line and do not require hospitalization. In contrast, cancers that are typically more aggressive or more likely to be diagnosed and cared for in the hospital setting (eg, lung and pancreatic cancer) are less likely to be recorded in primary care data.[15] Cancers that have broad definitions (ie, head and neck cancer) had wider variations in IRs across data sources, likely due to variations in the different coding systems.

The CR has been regarded as a gold standard for source data[16]; however, the completeness of data in the CR can vary based on the cancer type and calendar time.[14,23] The CR was created in 2013 by merging eight regional English registries into a single national registry.[30] Data collected by the CR come from multiple sources, resulting in inconsistencies in the incoming data.[23] A study comparing the correctness and completeness of breast cancer diagnosis records in CPRD Aurum and GOLD against the HES (2004–2019) and CR databases (2004–2016) reported that among the data linked to all sources, some (CPRD Aurum: 11.5%; CPRD GOLD: 12.1%) cases of malignant breast cancer recorded in the CPRD or HES databases were missing from the CR database.[14] There have been multiple initiatives in the UK to improve cancer recording in primary care data. In the 2000s, the "2-week wait" referral pathway was introduced to ensure that patients with symptoms indicative of cancer were seen by a specialist within 2 weeks of primary care referral.[31] In 2004, the Quality and Outcomes Framework was introduced as a voluntary incentive program for primary care practices in England with the aim of improving patient care by rewarding practices for achieving quality indicators across various areas, including cancer diagnosis.[32] Both of these initiatives may have improved cancer diagnosis recording in primary care.[28]

In our study, unadjusted survival times were similar between CPRD Aurum and GOLD, and our study demonstrated the impact of having the ONS cause-of-death information on survival time, where we found that survival probabilities for all cancers over time were higher with cancer-related deaths compared with any cause of death. The survival probabilities for more aggressive cancer types, where survival times are often short, were comparable for cancer-related death and any cause of death. Thus, ONS cause-of-death information may be less relevant for more aggressive cancer types.

The inclusion of data from March to April 2020, which coincided with the initial peak of the COVID-19 pandemic in the UK, warrants careful consideration. During this period, there was a marked increase in mortality rates and a significant strain on healthcare resources that influenced survival outcomes in patients with cancer. The pandemic led to disruptions in cancer screening, diagnosis, and treatment services, which may have introduced bias into survival estimates for this time frame. Nevertheless, the decision to include these data was made to ensure completeness and to reflect real-world conditions. Recent literature emphasizes that care should be taken when interpreting cancer survival estimates during the pandemic, as observed differences may be attributable both to genuine changes in outcomes and artefacts arising from healthcare disruptions.[33]

There were limitations to this study. First, the study period ended in 2018, due to the availability of HES and CR data at the time of download; however, study findings remain relevant for researchers. Furthermore, since the denominator for IR calculations was total person-years for all patients captured by the CPRD Aurum or GOLD data for a given year, we were unable to censor a patient's time at risk when they had an event, potentially resulting in underestimation of IRs.

Despite missing medical codes for some specific cancers (eg, ALL, brain, breast, head and neck, prostate, thyroid), the estimated rates remained within the range reported by external sources. This study focused on identification of cancers but not on other data elements needed for cancer research questions. Similarly, survival depends on factors such as stage, histology, and treatment regimen, which were not evaluated. Overall, notable strengths of this study are the use of a large dataset to assess 19 different cancer types and the use of CPRD Aurum and GOLD with linkage to HES and CR to describe the data landscape of cancer diagnosis recording.

Differences in recording practices between primary and secondary care settings, as well as variations in coding systems, have important implications for the accuracy and completeness of cancer diagnosis data. Cancer diagnoses recorded in primary care databases such as CPRD Aurum and CPRD GOLD may include cases without histological confirmation, potentially introducing false positives compared with cancer registry data, which typically require histological or cytological verification. Miscoding, administrative errors, and changes in coding practices over time may further affect the reliability of diagnosis recording. These factors should be considered when interpreting IRs and comparing data sources. The process of patient linkage restricts the study population to individuals registered in England who are eligible for linkage to HES, CR, or ONS data. This limitation may reduce the generalizability of our findings to other regions of the UK or internationally. Furthermore, the demographic composition of CPRD-linked populations may not fully reflect the diversity of the UK population, particularly with respect to ethnic groups. Recent studies have demonstrated that IRs for oncology indications can vary substantially by ethnicity.[34] Researchers should therefore be cautious when generalizing findings from CPRD-linked datasets and consider the potential impact of demographic composition on observed IRs and outcomes.

While our findings are robust within the UK context, external benchmarking is currently limited. To improve generalizability, future work should incorporate comparisons with international datasets such as EUROCARE and SEER. These comparisons would help contextualize our findings within broader epidemiological trends and support the applicability of our results beyond the UK. Furthermore, differences in other data elements such as biomarkers and disease stage should be explored in future linkage studies.

Conclusions

This comparison of diagnosis recording for 19 cancer types highlights the variation in cancer recording across different data sources and represents the most comprehensive comparison to date. Cancer recording in CPRD Aurum and GOLD captures a high proportion of diagnoses of most cancer types; however, the findings demonstrate variation in cancer recording across different data sources. For cancers where capture in the primary care data is lower, linkage to HES and/or CR may be recommended. Inclusion of ONS cause-of-death information demonstrated that survival probabilities for all cancers over time were higher with cancer-related deaths than with any cause of death. This provides additional information regarding cancer-specific survival and highlights the need for careful consideration of data sources and survival endpoints in future studies. The findings of this study can guide researchers to select the most appropriate data source for their study question. For studies of breast, prostate, and lung cancers, CPRD Aurum or GOLD alone may be sufficient; however, linkage to HES and/or CR is recommended where a more complete case capture is required.

Abbreviations

ALL, Acute lymphoblastic leukemia; AML, Acute myeloid leukemia; CPRD, Clinical Practice Research Datalink; CR, Cancer Registry; GP, General practitioner; HES, Hospital Episode Statistics; ICD-10, International Classification of Diseases 10th Revision; IR, Incidence rate; MM, Multiple myeloma; ONS, Office for National Statistics; UK, United Kingdom.

Data Sharing Statement

The datasets used during this study were sourced via the Clinical Practice Research Datalink, which are not publicly available, but the data analysis outputs may be available from the corresponding author upon reasonable request.

Ethics Approval and Informed Consent

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 National Cancer Registration and Analysis Service (NCRAS) data, copyright © (2018), reused with the permission of the Health and Social Care Information Centre. All rights reserved. This study was approved by Research Data Governance (22-00015).

Acknowledgments

Medical writing and editorial assistance were provided by Adivitiya Bihagara (PhD), Melanie Francis (MSc), Chrysi Petraki (PhD), and Daria Renshaw (BA) of IQVIA, which was funded by Amgen Ltd.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was funded by Amgen Inc.

Disclosure

Anna B Chaplin, Olia Archangelidi, David Neasham, and George Kafatos are employed by Amgen Ltd and own shares in Amgen Inc. Dr Anna B Chaplin reports non-financial support from IQVIA during the conduct of the stud. Ms Katrina Wilcox Hagberg reports grants from Amgen Ltd, grants from Amgen Inc, grants from BeiGene, grants from The Gerber Foundation, outside the submitted work; also, she is an unpaid volunteer member of the Clinical Practice Research Datalink Data Quality Advisory Group. The authors report no other conflicts of interest in this work.

References

1. Medicines and Healthcare products Regulatory Agency. Clinical practice research datalink. Available from: https://www.cprd.com/. Accessed October 31, 2025.
2. Medicines and Healthcare products Regulatory Agency. Primary care data for public health research. Available from: https://www.cprd.com/primary-care-data-public-health-research. Accessed October 31, 2025.
3. 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
4. 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
5. 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
6. 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
7. Medicines and Healthcare products Regulatory Agency. CPRD linked data. Available from: https://www.cprd.com/cprd-linked-data. Accessed October 31, 2025.
8. Jick SS, Kaye JA, Vasilakis-Scaramozza C, et al. Validity of the general practice research database. *Pharmacotherapy*. 2003;23(5):686–689. doi:10.1592/phco.23.5.686.32205
9. Thomas S, Edwards C, Smeeth L, Cooper C, Hall A. How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? *Arthritis Rheum*. 2008;59(9):1314–1321. doi:10.1002/art.24015
10. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the general practice research database: a systematic review. *Br J Clin Pharmacol*. 2010;69(1):4–14. doi:10.1111/j.1365-2125.2009.03537.x
11. 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(572):e128–e136. doi:10.3399/bjgp10X483562
12. 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

13. 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
14. 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
15. Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the general practice research database and linked cancer registries. *Pharmacoepidemiol Drug Saf*. 2013;22(2):168–175. doi:10.1002/pds.3374
16. 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
17. Somathilake G, Ford E, Armes J, et al. Evaluating the quality of prostate cancer diagnosis recording in CPRD GOLD and CPRD Aurum primary care databases for observational research: a study using linked English electronic health records. *Cancer Epidemiol*. 2025;94:102715. doi:10.1016/j.canep.2024.102715
18. Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer Epidemiol*. 2012;36(5):425–429. doi:10.1016/j.canep.2012.05.013
19. Rañopa M, Douglas I, van Staa T, et al. The identification of incident cancers in UK primary care databases: a systematic review. *Pharmacoepidemiol Drug Saf*. 2015;24(1):11–18. doi:10.1002/pds.3729
20. Jick S, 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
21. Medicines and Healthcare products Regulatory Agency. CPRD Aurum December 2024 (Version 2024.12.001). Accessed October 31, 2025. doi:10.48329/qfkt-kb64
22. Medicines and Healthcare products Regulatory Agency. CPRD GOLD December 2024 (Version 2024.12.001). Accessed October 31, 2025. doi:10.48329/se5g-a206
23. Henson KE, Elliss-Brookes L, Coupland VH, et al. Data resource profile: national cancer registration dataset in England. *Int J Epidemiol*. 2020;49(1):16–16h. doi:10.1093/ije/dyz076
24. NHS England. Hospital Episode Statistics (HES) and Office for National Statistics (ONS). linked mortality data guide; 2025. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data/hes-and-ons-linked-mortality-data-guide. Accessed October 31, 2025.
25. Cancer Research UK. Acute myeloid leukaemia (AML) incidence statistics; 2025. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-aml/incidence?_gl=1*xrmml*_gcl_au*MTcxODQ5NDQwNS4xNzUzMDg3MDMx*_ga*MTMwNTk3NTI2NC4xNzUzMDg3MDMx*_ga_58736Z2GNN*czE3NTMwODcwMzEkbzEkZzEkdDE3NTMwODcxMDEkajUyJGwwJGgw. Accessed October 31, 2025.
26. Cancer Research UK. Stomach cancer incidence statistics; Available from: 2025. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/stomach-cancer/incidence#heading-Two. Accessed October 31, 2025.
27. Cancer Research UK. Ovarian cancer incidence statistics; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer/incidence#heading-Two. Accessed October 31, 2025.
28. Whitfield E, White B, Barclay ME, et al. Differences in recording of cancer diagnosis between datasets in England: a population-based study of linked cancer registration, hospital, and primary care data. *Cancer Epidemiol*. 2025;94:102703. doi:10.1016/j.canep.2024.102703
29. 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
30. Di Girolamo C, Walters S, Benitez, 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 in 2013. *BMC Cancer*. 2018;18(1):492. doi:10.1186/s12885-018-4417-3
31. NHS Improvement. Ensuring better treatment: going further on cancer waits. Available from: https://www.england.nhs.uk/wp-content/uploads/sites/44/2017/11/Going-Further-on-Cancer-Waits.pdf. Accessed October 31, 2025.
32. NHS England. Quality and outcomes framework. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof. Accessed October 31, 2025.
33. Stannard R, Lambert PC, Lyratzopoulos G, Andersson TM, Khan S, Rutherford MJ. The long-lasting impacts of the COVID-19 pandemic on population-based cancer survival: what are the implications for data analysis? *Br J Cancer*. 2025;132(8):673–678. doi:10.1038/s41416-024-02931-0
34. Delon C, Brown KF, Payne NWS, Kotrotsios Y, Vernon S, Shelton J. Differences in cancer incidence by broad ethnic group in England, 2013-2017. *Br J Cancer*. 2022;126(12):1765–1773. doi:10.1038/s41416-022-01718-5