International Journal of Epidemiology, 2026, **55(2)**, dyag016
https://doi.org/10.1093/ije/dyag016
Data Resource Profile

IEA International Epidemiological Association
OXFORD

Data Resource Profile: Prostate cancer data from Clinical Practice Research Datalink linked hospital records, mortality data and cancer registry standardized to the Observational Medical Outcomes Partnership common data model (CPRD-PCa-OMOP)

Eng Hooi Tan[1], Danielle Newby[1], Daniel Prieto-Alhambra[1,2], Mandickel Kamtengeni[1], Antonella Delmestri[1,*], OPTIMA Consortium[†]

[1]Health Data Sciences, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom
[2]Department of Medical Informatics, Erasmus University Medical Centre, Rotterdam, The Netherlands
*Corresponding author. Nuffield Department of Orthopaedics, BOTNAR Research Centre, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7LD, United Kingdom. E-mail: antonella.delmestri@ndorms.ox.ac.uk
[†]See Acknowledgments section for a complete list of authors who are part of the OPTIMA Consortium as well as their affiliations.

Keywords CPRD, National Cancer Registration and Analysis Service, NCRAS, OMOP CDM, oncology, data harmonization, data linkage

Key Features

- CPRD-PCa-OMOP comprises men from England with incident prostate cancer (PCa) selected from the Clinical Practice Research Datalink (CPRD) GOLD and Aurum primary care databases, linked to hospital admissions, Office for National Statistics mortality data, and national cancer registry records.
- Each dataset was standardized to the Observational Medical Outcomes Partnership Common Data Model to facilitate data integration, collaborative research, external validation, and network studies. All the standardized datasets were merged to create CPRD-PCa-OMOP.
- Men aged ≥18 years, diagnosed with their first PCa in primary care between 2010 and 2022, and confirmed in the cancer registry, with ≥1 year of general practitioner registration, were included. There are 69 109 patients from 1611 English practices: 5566 from CPRD GOLD and 63 543 from CPRD Aurum.
- CPRD-PCa-OMOP includes information on demographics, diagnoses, medications, procedures, referrals, test results, hospitalizations, episodes of care, cancer diagnoses, tumour characteristics such as Tumour, Node, Metastasis staging and Gleason score, date of death, and primary cause of death.
- This comprehensive data resource will enable the study of risk factors, treatment patterns, and longitudinal outcomes in PCa. CPRD data access can be requested at https://www.cprd.com/access-data.

Data resource basics

Observational data for cancer patients in the UK

The digital collection of UK patient care data was started in the late 1980s to manage general practitioner (GP) practices and patients, and was expanded over the years to include hospital events and registry records, generating large data sources that are now used for observational research [1].

Observational data for cancer patients are complex, distributed across different care settings. This requires the integration of various pieces of information, such as tumour histology, biomarkers, and treatment regimens, as well as longitudinal comorbidities and comedications. In non-cancer observational research, cohorts can be defined by the absence or presence of a health condition. The additional challenges of conducting observational research in cancer arise from diagnostic attributes (cancer staging, histology, grade, and biomarkers) that can determine the treatment options and prognoses [2].

Clinical Practice Research Datalink

One of the UK data providers that allow cancer data integration through a linked data service is the Clinical Practice Research Datalink (CPRD). CPRD is a not-for-profit, cost-recovery UK government research service delivered by the Medicines and Healthcare products Regulatory Agency (MHRA) with support

Received: 20 June 2025. **Accepted:** 20 January 2026

from the National Institute for Health and Care Research, as part of the Department of Health and Social Care [3]. CPRD hosts two primary care observational databases: CPRD GOLD [4] and CPRD Aurum [5], hereinafter called GOLD and Aurum, respectively. GOLD contains records collected by using the Vision® software from historical contributing practices in England, and from historical and current contributing practices in Wales, Scotland, and Northern Ireland. Aurum contains records collected by using the EMIS Web® software from historical and current contributing practices in England only. CPRD can provide a variety of linked patient-level data from England for both GOLD and Aurum, including the Hospital Episode Statistics (HES) Admitted Patient Care (APC), Office for National Statistics (ONS) Mortality, and National Cancer Registration and Analysis Service (NCRAS) datasets [6]. These datasets have been utilized individually or as part of a linkage in numerous peer-reviewed publications [7].

CPRD-PCa-OMOP

CPRD-PCa-OMOP is a new data source created to characterize and assess long-term outcomes in patients with prostate cancer (PCa) in the context of the public–private partnership OPTIMA (Optimal Treatment for Patients with Solid Tumours in Europe Through Artificial Intelligence). The OPTIMA consortium aims to achieve an interoperable real-world data and evidence generation platform to improve care for patients with prostate, breast, and lung cancer. To create this data source, we used the GOLD [4, 8] and Aurum [5, 9] December 2023 releases, both linked to the latest available HES APC [10–12], NCRAS cancer registration (NCRASCR) [13–15], and ONS Mortality datasets [16–18].

To avoid possible patient duplications, we have excluded from GOLD those practices that migrated from the Vision® software to the EMIS Web® software within CPRD.

Because of the different data structures of these datasets and to facilitate the interoperability and execution of distributed network studies, we transferred each of these datasets to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5.4 [19] and merged the resulting databases. The OMOP CDM is a well-established standardization method, which has been adopted to facilitate interoperability and federated analytics in various frameworks such as the European Health Data and Evidence Network (EHDEN) consortium [20], the Data Analysis and Real World Interrogation Network (DARWIN) EU® initiative [21], and the OPTIMA consortium [22].

CPRD-PCa-OMOP is the first data source in OMOP CDM format that includes CPRD primary care data linked to HES APC, ONS Mortality, and NCRASCR datasets.

Data collected

CPRD-PCa-OMOP includes 69 109 men from England aged ≥18 years diagnosed with their first PCa in GOLD or Aurum between 2010 and 2022 and confirmed in NCRASCR, with ≥1 year of registration at a GP practice [23]. In CPRD, only consenting GP practices from England are eligible for linkage and we linked primary care records to HES APC, ONS Mortality, and NCRASCR data (Table 1). Patients who opt out of having their data utilized in clinical research were excluded from CPRD database releases. We excluded ONS Mortality- and HES-linked data that were provided with a weak linkage likelihood (1 to 5 = the most likely, the least likely), following a previously published algorithm [24]. We only accepted records with a linkage likelihood of between 1 and 2 (i.e. hes_patient.match_rank <=2). To prevent data duplication, we retained only the most recently registered patient identifier when multiple identifiers were linked to the same individual via a general identifier provided in the HES data (i.e. hes_patient.hes_gen_id).

Table 2 reports the patient demographics of CPRD-PCa-OMOP. There were 5566 patients in GOLD and 63 543 in Aurum. The age distribution was similar between both data sources, with most patients (38.40%) first diagnosed between 70 and 79 years of age. The median number of follow-up years from date of diagnosis until the end of observation was 6.13 [interquartile range (IQR) 3.71–8.68].

In Table 3, we describe the patient characteristics of CPRD-PCa-OMOP in terms of cancer stage and grade, and the prevalence of comorbidities and comedications. Most patients had Stage 1 cancer (26.16%). A higher proportion of patients had Stage 3 and 4 cancers in Aurum as compared with GOLD. Prevalent comorbidities before PCa diagnosis included hypertension (52.30%), hyperlipidaemia (24.77%), and type 2 diabetes (14.92%). Common medications prescribed in the year before PCa diagnosis included lipid-modifying agents (45.62%), drugs for acid-related disorders (33.68%), and drugs for obstructive airway diseases (20.34%).

Data standardization to the OMOP CDM

We standardized all the datasets described above to the OMOP CDM v. 5.4, the latest version recommended at the time of writing, to enable research collaboration through standardized analytical packages across database partners using a distributed federated network strategy. Although other CDMs exist, the

Table 1 Key details about CPRD-PCa-OMOP.

Countries covered	England
Data origin	CPRD Aurum, CPRD GOLD, HES APC, ONS Mortality, NCRAS
Who is included?	69 109 male patients with first PCa between 2010 and 2022 from 1611 practices (1420 in CPRD Aurum, 191 in CPRD GOLD, 18.33% of all practices from England)
What is recorded?	Demographics, diagnoses, symptoms, prescriptions, referrals, test results, hospitalizations and episodes of care, cancer diagnoses and cancer treatments, date of death, and primary cause of death
Start and end dates	From January 1995 to November 2023, with a median follow-up of 6.12 years from PCa diagnosis (IQR: 3.71–8.68)

IQR, interquartile range.

Table 2 CPRD-PCa-OMOP patient demographics.

Demographics	GOLD	Aurum	All
Number of patients	5566	63 543	69 109
Age at diagnosis (years) [*n* (%)]			
18–49	45 (0.81)	599 (0.94)	644 (0.93)
50–59	511 (9.18)	6330 (9.96)	6841(9.90)
60–69	1848 (33.20)	20 647 (32.49)	22 495 (32.55)
70–79	2082 (37.41)	24 454 (38.48)	26 536 (38.40)
80+	1080 (19.40)	11 513 (18.12)	12 593 (18.22)
Ethnicity [*n* (%)]			
Asian	12 (0.22)	253 (0.40)	265 (0.38)
Asian Indian/ Bangladeshi/Pakistani	46 (0.83)	773 (1.22)	819 (1.19)
Black	99 (1.78)	2463 (3.88)	2562 (3.71)
Chinese	8 (0.14)	89 (0.14)	97 (0.14)
White	5265 (94.59)	57 522 (90.52)	62 787 (90.85)
No matching concept	136 (2.44)	2443 (3.84)	2579 (3.73)
Follow-up			
Follow-up from diagnosis (years) [median (IQR)]	4.92 (2.65–7.14)	6.24 (3.88–8.83)	6.13 (3.71–8.68)
Death [*n* (%)]			
Death (all-cause)	1979 (35.56)	23 251 (36.59)	25 230 (36.51)
Death (PCa)	874 (15.70)	9290 (14.62)	10 164 (14.71)

IQR, interquartile range.

Table 3 CPRD-PCa-OMOP patient clinical characteristics.

Characteristics	GOLD patients	Aurum patients	All patients
Number of patients	5566	63 543	69 109
Cancer stage [*n* (%)]			
1	1430 (25.69)	16 650 (26.20)	18 080 (26.16)
2	982 (17.64)	10 248 (16.13)	11 230 (16.25)
3	816 (14.66)	11 207 (17.64)	12 023 (17.40)
4	831 (14.93)	10 769 (16.95)	11 600 (16.79)
Missing	1507 (27.08)	14 669 (23.09)	16 176 (23.41)
Gleason combined grade [*n* (%)]			
Low (2–6)	1210 (21.74)	12 707 (20.00)	13 917 (20.14)
Medium (7)	2225 (39.97)	25 407 (39.98)	27 632 (39.98)
High (8–10)	1204 (21.63)	15 083 (23.74)	16 287 (23.57)
Invalid[a]/missing	927 (16.65)	10 346 (16.28)	11 273 (16.31)
Charlson Comorbidity Index (2 years before PCa diagnosis) [*n* (%)]			
0	4764 (85.59)	52 939 (83.31)	57 703 (83.50)
1	454 (8.16)	5702 (8.97)	6156 (8.91)
2	178 (3.20)	2649 (4.17)	2827 (4.09)
≥3	144 (2.59)	2082 (3.28)	2226 (3.22)
Missing	26 (0.47)	171 (0.27)	197 (0.29)
Comorbidities any time before PCa diagnosis [*n* (%)]			
Hypertension	2563 (46.05)	33 582 (52.85)	36 145 (52.30)
Hyperlipidaemia	1301 (23.37)	15 815 (24.89)	17 116 (24.77)
Type 2 diabetes	760 (13.65)	9554 (15.04)	10 314 (14.92)
Comedications in the year before PCa diagnosis [*n* (%)]			
Lipid-modifying agents	2477 (44.50)	29 049 (45.72)	31 526 (45.62)
Drugs for acid-related disorders	1851 (33.26)	21 428 (33.72)	23 279 (33.68)
Drugs for obstructive airway diseases	1140 (20.48)	12 919 (20.33)	14 059 (20.34)

[a] Invalid values were any values outside of 2–10.

OMOP CDM has been chosen by health regulatory agencies to support policy-making [21] and is widely used nationally [3, 25] and internationally [26]. The harmonization of the dictionaries used in the datasets was produced by the EHDEN and Observational Health Data Sciences and Informatics (OHDSI) initiatives: standardized vocabularies are available for downloading at https://athena.ohdsi.org. For this study, we used the Athena vocabulary bundle v5.0 29-FEB-24 and the sources for standard mapped values are documented in Supplementary Tables S1–S11.

The data standardization to the OMOP CDM for primary care, secondary care, and mortality data was performed by following previously used and established techniques (https://github.com/oxford-pharmacoepi/etl_ndorms). For the cancer registration data, we developed new coding techniques by following the approach recommended by the OHDSI Oncology Working Group [2]. This approach, which provided part of the supporting rationale for the development of the OMOP CDM version 5.4, recognizes that cancer observational data are more challenging than those for most other conditions because they contain an individual's sequence of cancer states and treatments that define the journeys of patients through the disease. Because CPRD-PCa-OMOP includes ONS Mortality data, which are based on death certificates and are the primary source of information for NCRAS mortality, we did not utilize the NCRAS information on death. For all the datasets, source data field names and values were stored in the corresponding source reference variables in the CDM tables to maintain data provenance. Table 4 summarizes the information available in the linked datasets and the corresponding source and target vocabularies.

Cancer diagnoses

To map NCRAS cancer diagnoses of the 'tumour' dataset, we employed the World Health Organization International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) international standard for tumour registries, which uses a concatenation of cancer's attributes: [Histology]/[behaviour]-[Topography] (Supplementary Table S1). We mapped 69 106 PCa diagnoses to the OMOP CDM CONDITION_OCCURRENCE table by using 26 standard concept identifiers, of which 24 belonged to the ICD-O-3 vocabulary. One of the two standard concept identifiers that belonged to the Systemized Nomenclature of Medicine (SNOMED) vocabulary (4161028 = '8140/3-C61.9' = 'Adenocarcinoma of prostate') covered 92.1% of the diagnoses. Six Aurum patients had diagnoses not present in the OHDSI vocabularies: the latter were stored in the OMOP CDM OBSERVATION table.

Cancer modifiers

Whenever possible, we mapped cancer modifiers of the 'tumour' dataset (i.e. stage, grade, size, number of local and regional lymph nodes tested and those classified as metastatic, and Charlson comorbidities indexes over 2 and 6 years) into the OMOP CDM MEASUREMENT table by using the OHDSI 'Cancer Modifier' and SNOMED standardized vocabularies, for a total of 510 359 records. When cancer modifiers were not adequately represented in the OHDSI vocabularies, we stored them in the OMOP CDM OBSERVATION table, following the OHDSI best practice recommendations [27].

Cancer staging

Using the 'Cancer Modifier' vocabulary, we mapped several cancer stage attributes reported in the NCRAS dataset and classified them by using the American Joint Committee on Cancer (AJCC)/Union for International Cancer Control (UICC) Tumour, Node, Metastasis (TNM) system [28]. AJCC/TNM is a globally recognized standard that categorizes malignant tumours based on the primary tumour (T), the regional lymph node involvement (N), and distant metastasis (M). NCRAS derived stage attributes by combining related variables: 'stage_best' ('t_best', 'n_best', 'm_best'), 'stage_img' ('t_img', 'n_img', 'm_img'), and 'stage_path' ('t_path', 'n_path', 'm_path'). We used 43 standard concept identifiers to map stages 1[A–E] to 4[A–E] associated with different versions of the AJCC/UICC TNM staging system for different patients and diagnoses (i.e. 6th, 7th, 8th, and no version).

Cancer grading

We mapped to standard concept identifiers several cancer grade attributes: the 'Grade of tumour' (G1 to G4 = more abnormal cells, greater likelihood of aggressive growth and spread) (Supplementary Table S2) and the scoring system for malignant neoplasm of the prostate 'Gleason Primary/Secondary/Tertiary Pattern Grade' (1 to 5 = the more aggressive, the least differentiated) by using the 'Cancer Modifier' vocabulary (Supplementary Tables S3–S5), and the 'Gleason Combined Pattern Grade' (Supplementary Table S6) by using the SNOMED vocabulary.

There were 28 088 'Grade of tumour' instances, of which 51.08% were classified as G3, 32.94% as G2, 12.49% as G4, and 3.50% as G1.

The 57 852 'Gleason Combined' instances were associated with 21 distinct values, of which we accepted values from 2 to 10. The most frequent value was 7 (39.98%), followed by 6 (19.94%), 9 (13,36%), 8 (9.07%), and 10 (1.13%).

Cancer size

There were 440 instances of 'Size of tumour' reported as the diameter of a tumour in millimetres (if more than one tumour was present, the size of the largest tumour was reported). These instances were associated with 66 distinct values, of which the most frequent 5 (20, 15, 25, 10, and 12 mm) represented 33.18% of the total.

Cancer lymph nodes

There were 4482 instances of 'Number of local and regional metastatic lymph nodes'. Of these, 17.0% reported no positive lymph nodes (value = 0) and 50.4% reported values from 1 (8.41%) to 9 (3.10%). The remaining 32.62% covered values from 11 (3.10%) to 89 (0.02%).

There were 4375 instances of 'Number of local and regional lymph nodes examined' in excised specimens, with 89.74% reporting a value of 0.

Charlson comorbidities indexes

There were 68 912 instances of both Charlson comorbidity scores on diagnosis date: one covering 2 years, looking back from 27 to 3 months before diagnosis; and the other covering 6 years, looking back from 78 to 6 months before diagnosis. In both cases, the most frequent score was 0 (83.50% and 74.83%,

Table 4 Key information and vocabularies of CPRD-PCa-OMOP.

Dataset	Key information	Source vocabulary	OMOP table	Target vocabulary
CPRD	Diagnoses	Read SNOMED	condition_occurence	SNOMED, OMOP Extension
	Procedures	Read SNOMED	procedure_occurrence	SNOMED, CPT4, OMOP Extension
	Medications	dm+d gemscript	drug_exposure	RxNorm, RxNorm Extension CVX
	Medical devices	dm+d gemscript	device_exposure	dm+d, SNOMED
	Measurements	Read SNOMED	measurement	SNOMED, LOINC, OMOP Extension
	Anything else	Read SNOMED	observation	SNOMED + many
HES APC	Diagnoses	ICD-10	condition_occurence	SNOMED
	Procedures	OPCS-4	procedure_occurrence	SNOMED
ONS	Primary cause of death	ICD-10	death	SNOMED
NCRASCR	Cancer diagnoses	ICD-O-3 + many	condition_occurrence	SNOMED, ICD-O-3
	Cancer modifiers	Bespoken	measurement	Cancer Modifier, SNOMED
	Cancer procedures	OPCS-4	procedure_occurrence	SNOMED, OPCS-4
	Cancer medications	Bespoken RxNorm Extension	drug_exposure	RxNorm, RxNorm Extension

CPT4, Current Procedural Terminology, Fourth Edition; CVX, vaccine administered code; dm+d, NHS dictionary of medicines and devices; ICD-10, International Statistical Classification of Diseases and Related Health Problems, 10th Revision; ICD-O-3, International Classification of Diseases for Oncology, 3rd Edition; LOINC, Logical Observation Identifiers Names and Codes; NCRASCR, National Cancer Registration and Analysis Service cancer registration data; OPCS-4, Classification of Interventions and Procedures, version 4; SNOMED, Systemized Nomenclature of Medicine.

respectively), followed by 1 (8.91%, 13.49%), 2 (4.09%, 6.30%), 3 (1.80%, 2.80%), 4 (0.86%, 1.28%), and 5 (0.36%, 0.59%), with higher scores having minimum representation.

Cancer procedures

Cancer procedures from the NCRAS 'Treatment' dataset were mapped to the OMOP CDM PROCEDURE_OCCURRENCE table by using the OHDSI SNOMED and OPCS-4 standardized vocabularies (Supplementary Tables S7 and S8), for a total of 159 673 records. The five most frequent procedures accounted for 81.64% of the records, with 'Imaging' covering 40.07%, 'Hormone therapy' covering 20.27%, 'External beam radiation therapy procedure' covering 9.66%, 'Transrectal needle biopsy of prostate' covering 6.14%, and 'Radical prostatectomy' covering 5.50%.

Cancer treatments

Cancer treatments from the NCRAS 'Treatment' dataset were mapped to the OMOP CDM DRUG_EXPOSURE table by using OHDSI 'RxNorm' and 'RxNorm Extension' standardized vocabularies (Supplementary Table S9). A total of 2913 records were mapped using 39 standard concept identifiers and the five most frequent concepts covered 85.65% of the records. Bicalutamide accounted for 38.76%, goserelin for 26.16%, cyproterone for 14.07%, gonadorelin for 3.50%, and docetaxel for 3.16%.

Cancer episodes

All cancer diagnoses, modifiers, and treatments (Supplementary Table S10) were additionally mapped to the OMOP CDM EPISODE and EPISODE_EVENT tables to identify cancer patients' trajectories through the disease (Supplementary Table S11). A total of 243 095 episodes were created associated with the domains of Procedure (42.61%), Regimen (28.96%), and Condition (28.43%). The number of episode events was 1 548 516, linking the events to the OBSERVATION (46.40%), MEASUREMENT (38.64%), PROCEDURE_OCCURRENCE (10.31%), CONDITION_OCCURRENCE (4.46%), and DRUG_EXPOSURE (0.19%) tables.

Cancer deaths

All primary causes of death were mapped to the OMOP CDM DEATH table, together with the date of death. A total of 25 230 deaths were reported, with 40.29% (39.96% in Aurum, 44.16% in GOLD) associated with 'Primary malignant neoplasm of prostate' (concept_id = 200962).

Data-quality assessment

We assessed the quality of the data mapped to the OMOP CDM by using Kahn's framework [29], implemented by the OHDSI software DataQualityDashboard (DQD, version: 2.6.3), available at https://github.com/OHDSI/DataQualityDashboard. DQD performs validation and verification checks according to conformance, completeness, and plausibility at the concept, field, and table levels by using check-specific thresholds. A total of 4684 data-quality checks were performed on the data, of which 98% passed overall, with both conformance and completeness satisfying 99% of the quality assessment and plausibility 96%. The data-quality checks that failed were investigated and none of them was a cause for concern for our study (e.g. unit_concept_ids used in the UK were not recognized by DQD and a few checks contained software mistakes, all of which we reported to the DQD developers).

Data resource use

The linked datasets in their source format have been used in other studies [30] to investigate the quality and concordance of cancer diagnoses. CPRD-PCa-OMOP is a novel dataset mapped to the CDM to allow participation in multinational network cohort studies. The data resource was created to contribute data to the OPTIMA consortium and it will be used to answer questions prioritized by healthcare professionals/patient stakeholder groups, related to the characterization of patients with PCa, risk stratification, treatment patterns, and comparative effectiveness/safety analyses [31]. We are further investigating the characteristics and outcomes of patients who received prostatectomy versus radiotherapy in a separate study [32].

Strengths and weaknesses

The key strength of CPRD-PCa-OMOP is its data sources, which are comprehensive and representative of the population from England, with longitudinal follow-up information and linkage to hospital admissions, cancer registry data, and mortality data, the latter being particularly important for the accuracy of cancer survival prediction. Having linked data sources allows researchers to triangulate between data points and define the variables according to the study objectives. For example, the first date of cancer diagnosis varied between the different data sources. Whitfield *et al.* [33] found that agreement in diagnosis dates was variable across different cancer sites and showed higher agreement between HES and NCRAS. For the description of this data source, we used the date recorded in NCRAS as the reference standard. A small proportion of patients, however, had earlier dates in primary care and HES records. As noted by Whitfield *et al.* [33], earlier dates in primary care may reflect the coding of suspected cancers before confirmation. Similarly, earlier cancer diagnoses for a minority of patients in HES records may have resulted from delays in NCRAS recording. NCRAS records rely on pathological confirmation, which is a time-consuming process, whereas HES records capture the first clinically relevant diagnosis, such as one based on imaging. Another major strength of CPRD-PCa-OMOP is that it is OMOP-standardized, which facilitates data integration, collaborative research, external validation, and network studies. Mapping of cancer records has also been performed by Genomics England [34], but we have additionally mapped episodes to study patient trajectories and followed the OMOP CDM best practice [19].

The main weakness of CPRD-PCa-OMOP is the lack of information about chemotherapy and radiotherapy regimens, which we requested after obtaining ethical approval, but never received due to a delay in the release of the datasets held by a third party. This situation has limited the capacity of the study to investigate disease-free survival, relapse or recurrence, and cancer progression. However, we remain committed to standardizing the missing regimen data and releasing the corresponding GitHub code and documentation to the research community as soon as is feasible.

Some information regarding cancer stage (23.41%) and Gleason grade (16.31%), which are important prognostic factors in cancer research, was missing. However, Strongman *et al.* [30] showed that the completeness of cancer stage and grade has improved over time, especially since 2012.

Finally, data standardization could sometimes hinder interoperability between studies when different standard concepts are used to represent the same variables. To reduce this risk, we have provided detailed documentation of the standardization process along with the corresponding GitHub code. In addition, while concept harmonization is based on the evolving Athena vocabularies, their changes mainly include additions and error corrections, ensuring that the final outputs remain highly compatible.

Data resource access

Source data documentation are available at https://www.cprd.com/primary-care-data-public-health-research and https://www.cprd.com/cprd-linked-data. The Extract, Transform, and Load (ETL) process is documented at https://github.com/oxford-pharmacoepi/etl_ndorms and https://oxford-pharmacoepi.github.io/etl_ndorms/docs/NCRASCR. Researchers can apply for CPRD data access subject to licensing fees and protocol approval (enquiries: enquiries@cprd.com). For queries about the mapped dataset, please refer to the 'Supplementary Material' file, which provides detailed information on the standardized variables and on the harmonization choices. The data used in this study were provided under the University of Oxford multi-study CPRD licence and cannot be shared. However, the Python 3 code and the ETL documentation are freely available on GitHub and could be used to recreate or adapt the mapped database or as a template for other cancer data sources.

Ethics Approval

CPRD has ethics approval from the UK Health Research Authority to support research using anonymized patient data and must complete an annual NHS Data Security and Protection Toolkit assessment to demonstrate that it meets the required standard for holding data securely. CPRD data are accessible only via a CPRD client approval, a CPRD licence, and a protocol ethical approval obtained via the CPRD's Research Data

Governance process. CPRD data cannot be shared with collaborators not named in the protocol and they need to be erased when the protocol research question has been answered. Access to CPRD data and to the linkages used to create CPRD-PCa-OMOP was granted by protocol number 22_001867.

Acknowledgements

We thank Teen, Wai Yi Man, from the University of Oxford for her contribution to the data-conversion process into the OMOP CDM. This study is based in part on data from the CPRD obtained under license from the UK MHRA. 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 author/s alone.

The OPTIMA consortium (with associations): James N'Dow[1], Emma Jane Smith[1], Ellen Moyse[1], Carla Bezuidenhout[1], Torsten Gerriet Blum[1], Peter-Paul Willemse[1], Philip Cornford[1], Maurice Schlief[1], Alberto Briganti[7], Susan Evans Axelsson[1], Vasileios Sakalis[1], Anders Bjartell[2], Monique Roobol[3], Katharina Beyer[3], Lionne Venderbos[3], Sebastiaan Remmers[3], Giorgio Gandaglia[7], Raoul Boomsma[3], Carolina Testa[3], Charles Auffray[4], Jie Shen[36], Nareen Katta[36], Steven MacLennan[5], Sara MacLennan[5], Valerie Speirs[5], Yilin Xu[36], Imran Omar[5], Demi McDonald[5], Kelly Gray[5], Lesley Anderson[5], Charlotte Murray[5], Dianne Brown[5], Abdelali (Ali) Majdi[33], Carl Steinbeisser[33], Christoph Kowalski[22], Mart Kals[6], Nora Tabea Sibert[22], Juan Gómez Rivas[1], Martina Faticoni[7], Greta Matteuzzi[7], Stephane Lejeune[17], Louise Jones[8], Daan Nieboer[3], Javier Téllez[26], Guido Juckeland[9], Sandra Garrido[26], Daniel Kotik[9], Tobias Vu[9], Artur Yakimovich[9], Torsten Bauer[10], Jens Kollmeier[10], Roberto Galán[26], Ruben Villoria[26], Tobias Sjöblom[11], Chatarina Larsson[11], Ivaylo Stoimenov[11], Daniel Prieto-Alhambra[12], Sara Khalid[12], Enric Bousoño Borrull[26], Laura Tur Giménez[26], Mahkameh Mafi[12], Soralys Hernandez[26], Alvaro Morandeira Galban[26], Nathaly Thiel[13], Nikolaus Forgó[13], Antoni Napieralski[13], Martina Wimmer[13], Katharina Haimbuchner[13], Saskia Kaltenbrunner[13], Syed Shah[3], Kseniia Guliaeva[13], Torsten Blum[10], Rebecca Graebig-Rancourt[10], Michael Bussmann[9], Kevin Joubel[18], Talita Duarte-Salles[20], Anna Palomar[20], Nadia Harbeck[15], Julian Koch[15], Irene Lopez[20], Julia Kasprzak[15], Giuseppe Curigliano[14], Sindhu Naidu[16], Beatrice Taurelli Salimbeni[14], Zoe Soo[6], Amyn Bhamani[16], Emanuela Ferraro[14], Mario Campone[18], Jean-Sebastien Frenel[18], Carmen Criscitiello[14], François Bocquet[18], Marion Laloue[18], Bérengère Denoël[18], Marie Moisy[18], Julien Berry[18], Alexandre Auffray[18], Frank Flammanc[18], Hyacinthe Boni[18], Clémence Gaborieau[18], Josephine Bertin[18], Philippe Lambin[19], Anshu Ankolekar[19], Chiara Corti[14], Alba Tor Roca[20], Elena Dal Zotto[14], Andreas Kremer[27], Valérie Vaccaro[21], Thomy Tonia[21], Céline Genton[21], Maria Quaranta[27], Loic Marx[27], Pippa Powell[23], Clare Williams[23], Manou Kooy[25], Nils Christian[27], Christian Seeling[15], Edward Burn[12], Antonella Delmestri[12], Inmaculada Perea Fernández[26], Danielle Newby[12], Cheryl Tan[12], Marcel Hartig[32], Juan Miguel Auñón García[26], José Carlos Barrios González[26], Lynn McRoy[32], Chad Barnett[32], Humma Khan[32], Claude Chelala[8], Maryam Abdollahyan[8], Christian Bauer[27], Mariana Pina[27], Romain Tching[27], Philipp Matthies[28], Corinna Zur Bonsen-Thomas[28], Larissa Tschetsch[28], Francisco Pinto[28], Heinz Vargas[28], Matthieu Blottière[29], Louise Duflot[29], David Vallas[29], Patrizia Torremante[30], Verena Von Scharfenberg[30], Nuno Azevedo[31], Alla Kolyban[31], Javier Cuadrado Corz[8], Waltraud Kantz[32], Frederic Kube[32], Amanda Matthews[32], Bhakti Arondekar[32], Bruno Gori[32], Hagen Krüger[32], Julia Ilinares[32], Keith Wilner[32], Lucile Serfass[32], Joao Mouta[35], Robert Miller[32], Lara Chayab[35], Milan Tsompanoglou[32], Karen Godbold[32], Stefan Langhammer[32], Anne Adams[32], Sebastian Boie[32], Florian Reis[32], Eugenia Schander[32], Andres Metspalu[6], Hessa Bakhtiar[32], Stefan Krautschneider[32], Neda Rajaeean[32], John-Edward Butler-Ransohoff[33], Santiago Villalba[33], Adrian Wolny[33], Lisa Schneider[33], Jaak Vilo[6], Raivo Kolde[6], Adrian Rousset[34], Ivo Cleuren[34], Sandra Eketorp Sylvan[34], Ellie Paintin[34], Monika Pokrzepa[34], Carolin Lorber[35], Stefanie Morris[35], Sulev Reisberg[6], Marek Oja[6], Camille Andre[35], Tobias Schulte in den Baeumen[35], Jason Hannon[35], Kartick Sukumaran[35], Telver Objartel[6], Neal Navani[16], Sam Janes[16], Sean C Turner[36], David Dellamonica[37], Heather Moses[37], Yiduo Zhang[37], Christophe Dufour[37], Marcus Simon[37], Hassan Naqvi[37], Jens Ceder[37], Olga Alekseeva[37], Burkhard Mueller[38], Tobias Flosdorf[38], Daniel Kraenz[38], Anastasia Goette[38], Gustaf Hedström[39], Peter Hellberg[39], Per-Henrik Edqvist[39], Rossella Nicoletti[40], Mauro Gacci[40], Francesco Annunziato[40], Bertrand De Meulder[41], Nadezhda Knauer[41], Djeanne Onthoni[41], Maxence Renaud[41], Miguel Taraud[41]

[1]Stichting European Urological Foundation, Arnhem, Netherlands
[2]Lunds Universitet, Lund, Sweden
[3]Erasmus Universitair Medisch Centrum Rotterdam, Rotterdam, Netherlands
[4]Association EISBM, Vourles, France
[5]The University Court of the University of Aberdeen, Aberdeen, United Kingdom
[6]Tartu Ulikool, Tartu, Estonia
[7]Universita Vita-Salute San Raffaele, Milan, Italy
[8]Queen Mary University of London, London, United Kingdom
[9]Helmholtz-Zentrum Dresden-Rossendorf EV, Dresden, Germany
[10]Helios Klinikum Emil von Behring GMBH, Berlin, Germany
[11]Uppsala Universitet, Uppsala, Sweden
[12]The Chancellor, Masters and Scholars of the University of Oxford, Oxford, United Kingdom
[13]Universitat Wien, Vienna, Austria
[14]Istituto Europeo Di Oncologia SRL, Milan, Italy
[15]Ludwig-Maximiliansuniversitaet Muenchen, Munich, Germany
[16]University College London, London, United Kingdom
[17]European Organisation for Research and Treatment of Cancer AISBL, Brussels, Belgium
[18]Institut De Cancerologie De L'ouest, Angers, France
[18]Universiteit Maastricht, Maastricht, Netherlands
[20]Fundacio Institut Universitari Pera La Recerca A L'atencio Primaria De Salut Jordi Gol I Gurina, Barcelona, Spain
[21]European Respiratory Society, Lausanne, Switzerland
[22]Deutsche Krebsgesellschaft Ev, Berlin, Germany
[23]European Lung Foundation, Sheffield, United Kingdom
[24–25]PNO Life Sciences and Health, Rijswijk, Netherlands
[26]GMV Soluciones Globales Internet SAU, Madrid, Spain
[27]Information Technology for Translational Medicine (ITTM) SA, Esch-sur-Alzette, Luxembourg

[28]Smart Reporting GMBH, Munich, Germany
[29]Owkin France, Paris, France
[30]PNO Innovation GMBH, Munich, Germany
[31]Ydeal.Net Software LDA, Santa Maria Da Feira, Portugal
[32]Pfizer Limited, Sandwich, United Kingdom
[33]Bayer Aktiengesellschaft, Leverkusen, Germany
[34]Amgen, Diegem, Belgium
[35]F. Hoffmann-la Roche AG, Basel, Switzerland
[36]Abbvie inc, North Chicago, United States
[37]Astrazeneca AB, Södertälje, Sweden
[38]Mutabor Technologies GMBH, Hamburg, Germany
[39] Region Uppsala, Uppsala, Sweden
[40]Universitá di Firenze, Florence, Italy
[41]BDM Consulting, Lyon, France

Author contributions

A.D. led the OMOP data standardization. A.D. and E.H.T. ran the descriptive statistics and wrote the paper. M.K. performed the data-quality checks. All the authors critically reviewed the manuscript. A.D. will act as guarantor for the paper.

Supplementary material

Supplementary material is available at *IJE* online.

Conflicts of interest

The Health Data Sciences research group has received research grants from the European Medicines Agency, from the Innovative Medicines Initiative, from Amgen, Chiesi, and from UCB Biopharma; as well as consultancy or speaker fees from Astellas, Amgen, AstraZeneca, and UCB Biopharma. E.H.T. received consultancy fees from Janssen Pharmaceutica NV, outside the submitted work. All other authors declare no additional conflicts of interest.

Funding

OPTIMA is funded through the IMI2 Joint Undertaking and is listed under grant agreement No. 101034347. IMI2 receives support from the European Union's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations (EFPIA). IMI supports collaborative research projects and builds networks of industrial and academic experts to boost pharmaceutical innovation in Europe. The views communicated within are those of OPTIMA. Neither the IMI nor the European Union, EFPIA, nor any Associated Partners are responsible for any use that may be made of the information contained herein. The interpretation and conclusions contained in this study are those of the author/s alone.

Data availability

The data underlying this article were provided by the CPRD under the multi-study licence held by the University of Oxford. These data can be requested at https://www.cprd.com/access-data by CPRD licence holders.

Use of Artificial Intelligence (AI) Tools

AI was not used in collecting and/or analysing the data, producing images or graphical elements, or in writing this paper.

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**:766–8. https://doi.org/10.1136/bmj.302.6779.766
2. Belenkaya R, Gurley MJ, Golozar A *et al.* Extending the OMOP common data model and standardized vocabularies to support observational cancer research. *JCO Clin Cancer Inform* 2021;**5**:12–20. https://doi.org/10.1200/cci.20.00079
3. Medicines and Healthcare products Regulatory Agency (MHRA). *Clinical Practice Research* Datalink *(CPRD)*. https://cprd.com/ (31 October 2025, date last accessed).
4. Sanchez-Santos MT, Axson EL, Dedman D, Delmestri A. Data Resource Profile Update: CPRD GOLD. *Int J Epidemiol* 2025;**54**:dyaf077. https://doi.org/10.1093/ije/dyaf077
5. Wolf A, Dedman D, Campbell J *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019;**48**:1740–g. https://doi.org/10.1093/ije/dyz034
6. Medicines and Healthcare products Regulatory Agency (MHRA). *CPRD linked data*. https://www.cprd.com/cprd-linked-data (31 October 2025, date last accessed).
7. Medicines and Healthcare products Regulatory Agency (MHRA). *CPRD Bibliography*. https://www.cprd.com/bibliography (31 October 2025, date last accessed).
8. Clinical Practice Research Datalink. *CPRD GOLD December 2023 dataset*. https://doi.org/10.48329/30pm-xq61 (31 October 2025, date last accessed).
9. Clinical Practice Research Datalink. *CPRD Aurum December 2023 dataset*. https://doi.org/10.48329/7njs-8a57 (31 October 2025, date last accessed).
10. Clinical Practice Research Datalink. *CPRD GOLD HES APC January 2022*. https://doi.org/10.48329/fagm-ez75 (31 October 2025, date last accessed).
11. Clinical Practice Research Datalink. *CPRD Aurum HES APC January 2022*. https://doi.org/10.48329/vagx-9d96 (31 October 2025, date last accessed).
12. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017;**46**:1093–i. https://doi.org/10.1093/ije/dyx015

13. Clinical Practice Research Datalink. *CPRD GOLD Cancer Registration August 2021*. https://doi.org/10.48329/541y-nh70 (31 October 2025, date last accessed).
14. Clinical Practice Research Datalink. *CPRD Aurum Cancer Registration August 2021*. https://doi.org/10.48329/5sm7-3209 (31 October 2025, date last accessed).
15. Henson KE, Elliss-Brookes L, Coupland VH *et al.* Data Resource Profile: National Cancer Registration Dataset in England. *Int J Epidemiol* 2020;**49**:16–h. https://doi.org/10.1093/ije/dyz076
16. Clinical Practice Research Datalink. *CPRD GOLD ONS deaths January 2022*. https://doi.org/10.48329/WQ7V-X832 (31 October 2025, date last accessed).
17. Clinical Practice Research Datalink. *CPRD Aurum ONS deaths data January 2022*. https://www.cprd.com/cprd-aurum-ons-deaths-data-january-2022 (31 October 2025, date last accessed).
18. Office for National Statistics (ONS). *User guide to mortality statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017 (31 October 2025, date last accessed).
19. Observational Health Data Sciences and Informatics (OHDSI). *OMOP Common Data Model (version 5.4)*. https://ohdsi.github.io/CommonDataModel/cdm54.html (31 October 2025, date last accessed).
20. European Health Data and Evidence Network (EHDEN). *EHDEN Consortium*. https://www.ehden.eu/consortium-partners/ (31 October 2025, date last accessed).
21. European Medicines Agency. *Data Analysis and Real World Interrogation Network (DARWIN EU)*. https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/real-world-evidence/data-analysis-real-world-interrogation-network-darwin-eu (31 October 2025, date last accessed).
22. OPTIMA. *OPTIMA Consortium*. https://www.optima-oncology.eu/optima-consortium/ (31 October 2025, date last accessed).
23. UK Parliament. *Constituency data: GPs and GP practices*. https://commonslibrary.parliament.uk/constituency-data-gps-and-gp-practices/ (31 October 2025, date last accessed).
24. Delmestri A, Prieto-Alhambra D. CPRD GOLD and linked ONS mortality records: Reconciling guidelines. *Int J Med Inform* 2020;**136**:104038. https://doi.org/10.1016/j.ijmedinf.2019.104038
25. NHS Digital. *The* NHS *Research Secure Data Environment Network*. https://digital.nhs.uk/data-and-information/research-powered-by-data/sde-network (31 October 2025, date last accessed).
26. European Medicines Agency. *Data Partners onboarded in DARWIN EU (Phase I II III)* https://www.ema.europa.eu/en/documents/other/darwin-eu-data-partners-onboarded-phases-i-ii-iii_en.pdf (31 October 2025, date last accessed).
27. Observational Health Data Sciences and Informatics (OHDSI). *OMOP Common Data Model (version 5.4) - Observation table*. https://ohdsi.github.io/CommonDataModel/cdm54.html#observation (31 October 2025, date last accessed).
28. Union for International Cancer Control (UICC). *TNM Classification of Malignant Tumours*. https://www.uicc.org/what-we-do/sharing-knowledge/tnm (31 October 2025, date last accessed).
29. Kahn MG, Callahan TJ, Barnard J *et al.* A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016;**4**:1244. https://doi.org/10.13063/2327-9214.1244
30. Strongman H, Williams R, Bhaskaran K. What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? a concordance and validation study using linked English electronic health records data. *BMJ Open* 2020;**10**:e037719. https://doi.org/10.1136/bmjopen-2020-037719
31. Omar MI, MacLennan S, Ribal MJ *et al.*; PIONEER Consortium. Unanswered questions in prostate cancer - findings of an international multi-stakeholder consensus by the PIONEER consortium. *Nat Rev Urol* 2023;**20**:494–501. https://doi.org/10.1038/s41585-023-00748-9
32. Tan E, Newby D, Man W *et al.* Comparison of baseline characteristics in patients with prostate cancer who received radical prostatectomy or external beam radiotherapy: target trial versus real-world data population. *Pharmacoepidemiol Drug Saf* 2025;**34**:130. https://doi.org/10.1002/pds.70186
33. 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. https://doi.org/10.1016/j.canep.2024.102703
34. Genomics England. *Public OMOP Mappings.* https://gitlab.com/genomicsengland/genomics_england_publications/public-omop-mappings (31 October 2025, date last accessed).

International Journal of Epidemiology, 2026, 55, 1–9
https://doi.org/10.1093/ije/dyag016
Data Resource Profile