International Epidemiological Association

International Journal of Epidemiology, 2019, 1740–1740g
doi: 10.1093/ije/dyz034
Advance Access Publication Date: 11 March 2019
Data Resource Profile

Data Resource Profile

Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum

Achim Wolf,* Daniel Dedman, Jennifer Campbell, Helen Booth, Darren Lunn, Jennifer Chapman and Puja Myles

Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, London, UK

*Corresponding author. Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, 10 South Colonnade, London E14 4PU, UK. E-mail: achim.wolf@mhra.gov.uk

Editorial decision 13 February 2019; Accepted 25 February 2019

Data resource basics

Clinical Practice Research Datalink (CPRD) is a UK government, not-for-profit research service that has been supplying anonymized primary care data for public health research for more than 30 years. In October 2017 CPRD launched a new data resource called CPRD Aurum. CPRD Aurum is a database containing routinely collected data from primary care practices in England, capturing diagnoses, symptoms, prescriptions, referrals and tests for over 19 million patients as of September 2018 (Figure 1). Primary care data in CPRD Aurum have been linked to national secondary care databases as well as deprivation and death registration data (Table 1).

UK primary care

The United Kingdom's (UK) National Health Service (NHS) is a publicly funded health service, free at the point of use. General practitioners (GPs) are considered the 'gatekeepers' of the NHS, referring patients to secondary care and diagnostic tests.[1] Over 98% of the population is registered at one of the 7300 GP practices in England.[2] A unique patient identifier, the NHS number, is used in primary, secondary and tertiary care settings, enabling linkages to other data sources.[3] There are four principal GP IT systems (primary care patient management software system) suppliers in England[4] and the largest coverage is provided by EMIS Health® (EMIS Web® software is used in 56% of English practices).[5] CPRD Aurum, discussed in this Data Resource Profile, encompasses EMIS Web® GP practices that have agreed to contribute data to this database on a daily basis. CPRD also collects data from practices using Vision® software that contribute to the CPRD GOLD database, which has been used in epidemiological research for 30 years.[6]

CPRD Aurum

CPRD Aurum includes patient electronic healthcare records (EHR) collected routinely in primary care. When a practice agrees to contribute patient data to CPRD Aurum, CPRD receives a full historic collection of the coded part of the practice's electronic health records, which includes data on deceased patients and those who have left the practice. Since 25 May 2018, individuals in England can opt-out of sharing their confidential patient information for research purposes[7] and, as of 1 September 2018, 2.7% of the English primary care registered population had opted-out.[8]

As of September 2018, CPRD Aurum included 7 million patients who were alive and registered at EMIS Web® currently contributing practices (Table 2), representing around 13% of the population of England. This number will increase as additional practices sign-up as contributors to this data resource as part of an ongoing recruitment strategy. Consenting practices from Northern Ireland will start contributing data to CPRD Aurum from 2019.

Key demographic information on current and total patients is presented in Table 2. Median follow-up since 1995 for all patients was 4.2 years [interquartile range (IQR): 1.5–11.4] and 9.1 years (3.3–20.1) for current patients. For patients in CPRD Aurum, the mean decile on the 2015 Index of Multiple Deprivation[9] was 5.3 (1 being the least deprived) compared with the mean decile in England (of 5.5), suggesting a slightly less deprived population in the database.

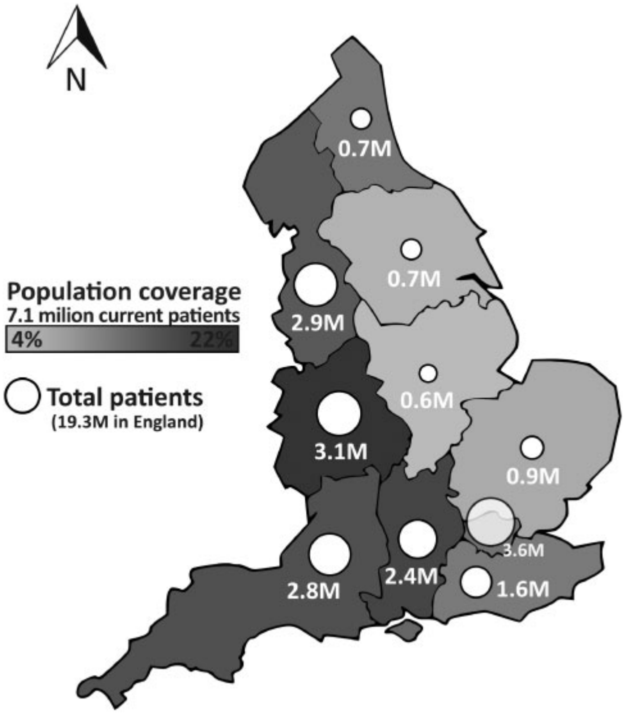

Figure 1. CPRD Aurum population coverage and total patients by English region, September 2018. Circles represent total patients in CPRD Aurum in each region. Shading represents population coverage of current patients as a proportion of total regional population.

Linkage to other datasets

Data from patients from all practices in CPRD Aurum can be linked to a range of health-related data sources including secondary care, disease registries and death registration records (Table 3). CPRD Aurum are linked to other patient-level health data by a trusted third party, NHS Digital, using NHS number, exact date of birth, sex and patient residence postcode (linkage methodology details are described in Padmanabhan *et al.*, 2018).[10] CPRD does not receive or hold patient identifiers including name, full date of birth, postcode and NHS number. Identifiers are removed prior to transfer of data to CPRD to protect patient confidentiality. Personal identifiers are sent separately from GP practices to NHS Digital, the statutory body in England able to receive patient identifiable data, to enable linkage.

Primary care data in CPRD Aurum have been linked to Office for National Statistics Death Registration Data,[11] which are considered the gold standard for mortality data in the UK and contain the date, place and cause of death.[12] The Hospital Episode Statistics (HES) datasets include Admitted Patient Care (APC) data which contain details of all admissions to, or attendances at English NHS health care providers, including acute hospital trusts, primary care trusts and mental health trusts.[13] HES Outpatient (OP) contains records of outpatient appointments in England including dates, specialty, clinical diagnoses and procedures.[14] HES Accident and Emergency (A&E) consists of individual records of patient care administered in the accident and emergency setting in England, also including diagnoses and procedures.[15] The HES Diagnostic Imaging Dataset (DID) contains information about diagnostic imaging tests conducted, such as X-rays and MRI scans, taken from NHS radiological information systems.[16]

Table 1. Key details about CPRD Aurum

UK countries covered	Consenting practices in England (consenting practices from Northern Ireland available from 2019)
Who is included?	19 million patients from 738 practices (10% of English practices), of whom 7 million were alive and currently contributing (13% of the population of England)
What is recorded?	Demographics, diagnoses, symptoms, prescriptions, referrals, immunizations, lifestyle factors, tests and results
Start and end dates	From 1995[a] to September 2018, with a median follow-up of 4.2 years (IQR: 1.5–11.4) for all patients and 9.1 years (IQR: 3.3–20.1) for current patients. CPRD Aurum is updated on a monthly basis.
Standard linkages	Hospital Episode Statistics, Death Registration, Cancer data, Mental Health Services Dataset, Small Area-Level Data (deprivation measures and rural–urban classification)

[a] This is an arbitrary cut-off and the database includes records pre-dating 1995, however, the completeness of recorded information following this point is expected to be more reliable.

IQR, interquartile range.

Table 2. Demographic characteristics of Aurum patients, September 2018

	All patients	Current
No. patients (practices)	19 305 234 (738)	7 125 786 (731)
Gender		
Male	9 309 928 (48.2%)	3 552 291 (49.9%)
Female	9 994 725 (51.8%)	3 573 360 (50.1%)
Indeterminate	581 (<0.01%)	135 (<0.01%)
Age in 2018		
<18	–	1 427 297 (20.0%)
18–64	–	4 463 385 (62.6%)
65+	–	1 235 104 (17.3%)
English Region		
North East	741 657 (3.8%)	327 786 (4.6%)
North West	2 948 404 (15.3%)	1 186 522 (16.7%)
Yorkshire & The Humber	715 487 (3.7%)	262 132 (3.7%)
East Midlands	551 130 (2.9%)	186 831 (2.6%)
West Midlands	3 126 234 (16.2%)	1 298 818 (18.2%)
East of England	889 553 (4.6%)	356 308 (5.0%)
South West	2 781 559 (14.4%)	980 184 (13.8%)
South Central	2 359 844 (12.2%)	847 481 (11.9%)
London	3 623 487 (18.8%)	1 125 905 (15.8%)
South East Coast	1 567 879 (8.1%)	553 819 (7.8%)
Follow-up since 1995[a] (median years, IQR)	4.2 (1.5–11.4)	9.1 (3.3–20.1)

'Current' refers to patients who are alive and registered at actively contributing practices. IQR, interquartile range.

[a]The database includes records pre-dating 1995.

Table 3. Standard linkages with CPRD Aurum data

Linkage dataset	Coverage[a]	Key information (including coding/scoring system)
ONS Death Registration Data	1998–2018	Date, place, and causes of death (ICD)
Hospital Episode Statistics (HES)		
Admitted Patient Care	1997–2017	Diagnoses (ICD) and procedures (OPCS)
Outpatient	2003–2017	Diagnoses (ICD) data
Accident & Emergency	2007–2017	Diagnoses (A&E codes) data
Diagnostic Imaging Dataset	2012–2017	Imaging tests data
PROM	2009–2017	Quality of life & condition-specific scales
National Cancer Registration and Analysis Service		
Cancer registration	1990–2015	Diagnoses (ICD) and tumour site
Systemic Anti-Cancer Treatment	2014–2015	Procedures and outcomes (ICD & OPCS)
National Radiotherapy Dataset	2012–2015	Procedures (ICD & OPCS)
Cancer Patient Experience Survey	2010–2013	Self-reported cancer patient data
Mental Health Services Data Set	2007–2015	Diagnoses (ICD), functioning (HoNOS)
Small area-level data		
Index of Multiple Deprivation	2004–2015	Patient or practice data, including domains
Townsend Index	2001	Patient-level deprivation data
Carstairs Index	2011	Practice-level deprivation data
Rural Urban Classification	2011	Practice-level classification

PROM, Patient Reported Outcome Measures; ICD, WHO International Classification of Diseases; OPCS, Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures; HoNOS, Health of the Nation Outcome Scale.

[a]Coverage is updated regularly based on data-provider releases. For up to date information on available linkages, visit cprd.com/linked-data

Cancer data provided by Public Health England (PHE) via the National Cancer Registration and Analysis Service (NCRAS)[17] have also been linked to CPRD Aurum. Linked NCRAS CPRD datasets include Cancer Registration data (record for each registrable tumour diagnosed or treated in England), the Systemic Anti-Cancer Treatment Dataset (SACT; chemotherapy treatment and outcome), the National Radiotherapy Dataset (RTDS;

radiotherapy records for cancer, including teletherapy and brachytherapy) and the Cancer Patient Experience Survey (CPES; four waves of self-reported patient data).

CPRD Aurum data has also been linked to the Mental Health Services Dataset (MHDS), which contains records of individuals who accessed secondary care adult, and child and adolescent mental health services, including diagnoses and episodes of care. Small area-level linkages on practice or patients' residence postcodes include several measures of area-level deprivation (Index of Multiple Deprivation,[9] Townsend Index,[18] and Carstairs Index[19]) and practice-level rural–urban classification.[20]

Data collected

CPRD Aurum is a dynamic database, and data are collected from contributing practices on a daily basis and processed to create monthly snapshots for observational research. This full de-identified coded clinical record includes symptoms, diagnoses, prescriptions, immunizations, tests, lifestyle factors and referrals recorded by the GP or other practice staff, but does not include free text medical notes.

Structure

The database structure is based on eight separate files, each containing patients' pseudonymized identifiers (Figure 2). The patient file records basic patient demographics, date of death if applicable and details on when the patient registered/deregistered from the practice. The practice file contains the practice region and the most recent date of data collection for the practice, and the staff file contains the job category for each staff member in CPRD Aurum. The consultation file holds information relating to the type of consultation as entered by the GP (e.g. telephone, home visit, practice visit), which can be internally linked (within CPRD Aurum) to observations that occur during the consultation via the consultation identifier, and to the staff member that conducted the consultation via the staff file.

The observation file contains the medical-history data entered on the GP system including symptoms, clinical measurements, laboratory test results, and diagnoses, as well as demographic information recorded as a clinical code (e.g. patient ethnicity). Observations that occur during a consultation are linked via the consultation identifier. CPRD Aurum data are structured in a long format (multiple rows per subject), and observations are linked to a parent observation. For example, measurements of systolic and diastolic blood pressure will be grouped together via a parent observation for blood pressure measurement.

Data in the referral and problem files are linked to the observation file and contain 'add-on' data for referral-type and problem-type observations. The referral file contains information involving both inbound and outbound patient referrals to or from external care centres (most frequently from the practice to a secondary care provider). The problem file contains details of the patient's medical history that have been defined by the GP as a 'problem', including the significance of the problem and its expected duration. GPs may use 'problems' to manage chronic conditions, thus enabling them to group clinical events (including drug prescriptions, measurements and symptoms) by problem rather than chronologically by consultation date. Finally,

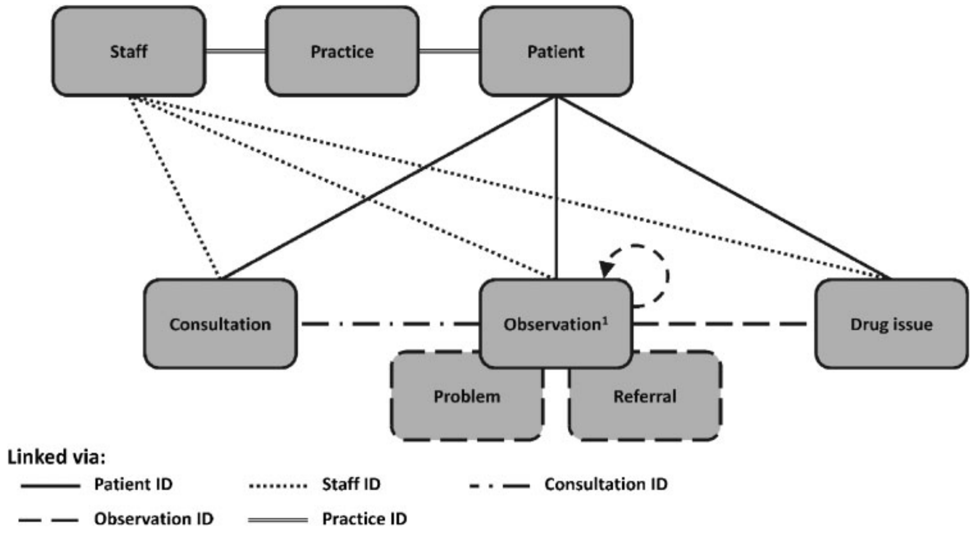

Figure 2. CPRD Aurum dataset structure. ¹Includes symptoms, diagnoses, immunizations, tests and lifestyle factors. *Note*: the problem and referral tables contain add-on information for certain types of observations. Some consultations are linked to observations. Some drug issues are linked to problem-type observations.

the drug-issue file contains data relating to all prescriptions (for drugs and devices) issued by the GP and are linked back to problem-type observations.

Coding

CPRD provides data dictionaries and code browsers to identify relevant codes in CPRD Aurum. The Medical Dictionary contains information on all medical history observations that have been recorded. Observations are coded using a combination of SNOMED CT (UK edition),[21] Read Version 2[22] and local EMIS Web® codes. The Drug Dictionary contains information on drug and device prescriptions recorded in EMIS Web®. This information is coded using the Dictionary of Medicines and Devices (dm+d), which exists within the SNOMED CT terminological structure.[23]

Practice staff are able to add additional information to patient records as free text. However, for data governance reasons, CPRD does not collect free text as these fields may contain identifiable patient information.

Ethics

CPRD obtains annual research ethics approval from the UK's Health Research Authority (HRA) Research Ethics Committee (REC) (East Midlands – Derby, REC reference number 05/MRE04/87) to receive and supply patient data for public health research. Therefore, no additional ethics approval is required for observational studies using CPRD Aurum data for public health research, subject to individual research protocols meeting CPRD data governance requirements.

Funding

CPRD is jointly sponsored by the UK government's Medicines and Healthcare products Regulatory Agency and the National Institute for Health Research (NIHR). As a not-for-profit UK government body, CPRD seeks to recoup the cost of delivering its research services to academic, industry and government researchers through research user licence fees.

Data resource use

As CPRD Aurum is a relatively recent data resource, no studies have been published to date. Nevertheless, since the recent launch of CPRD Aurum a number of approved research projects are already underway – including in pharmacovigilance, drug prescribing patterns, health services and policy evaluation, and disease risk factors. For instance, an ongoing academic research study is looking at patient treatment pathways in primary and secondary care using linked hospital and clinical audit data. A pharmacovigilance study is using CPRD Aurum and linked hospital data to examine the association between common drug therapies and heart arrhythmia. CPRD GOLD, which contains comparable NHS GP data to CPRD Aurum (but based on practices using a different GP IT system) has been used extensively in over 2000 publications that illustrate the potential research applications of CPRD Aurum.[24] A bibliography of all peer-reviewed published studies using CPRD data, dating back over the past 30 years, is available on the CPRD website (www.cprd.com/bibliography).

Strengths and weaknesses

With records on over 19 million patients as of September 2018, CPRD Aurum contains a wide range of diagnostic, prescription, procedure and lifestyle information. The key strengths of CPRD Aurum are its size and coverage, longitudinal follow-up, representativeness, standard linkages and data quality assurance processes.

Strengths

Database size and representativeness

The data currently cover 13% of the population of England, and are representative of the broader English population in terms of geographical spread (Figure 1) and deprivation (median decile on index of multiple deprivation (IMD) of 5.3), as well as age and gender [see Figure 3 comparing mid-2017 CPRD Aurum to mid-2017 data published by the Office for National Statistics (ONS)].[25]

Data linkages

Patient-level data have been linked to secondary care and other data sets, providing a fuller picture of the patient care pathway and outcomes. All CPRD Aurum practices have consented to participating in the linkage scheme, which includes data from national secondary care databases (hospitals and mental health service providers), the national cancer registry, death registrations and deprivation measures (Table 3).

Data quality assurance processes

CPRD undertakes various levels of validation and quality assurance on the daily GP data collection comprising over 900 checks covering the integrity, structure and format of the data. Issues highlighted by the checks are reviewed and addressed before data is incorporated into CPRD Aurum.

Collection-level validation ensures integrity by checking that data received from EMIS Web® practices contain only

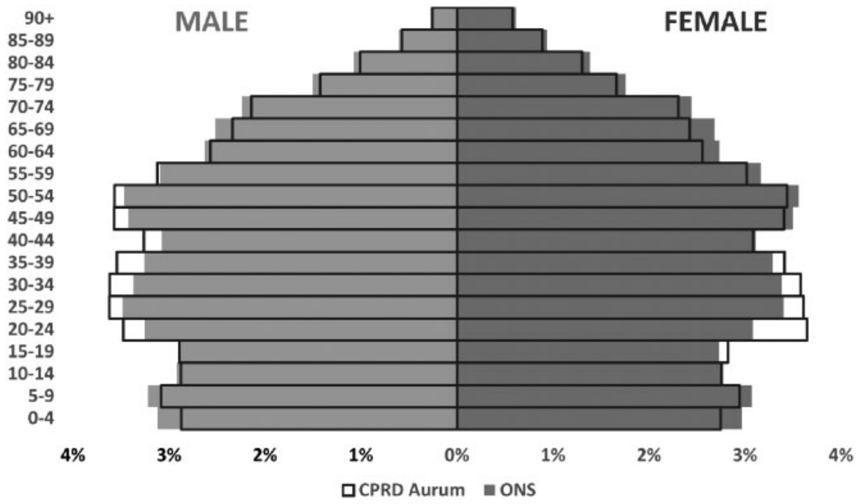

Figure 3. Population pyramids for CPRD Aurum and ONS data. Based on mid-2017 ONS and mid-2017 CPRD Aurum data.

expected data files and ensures that all data elements are of the correct type, length and format. Duplicate records are identified and removed. Transformation-level validation checks for referential integrity between records to ensure that there are no orphan records included in CPRD Aurum (e.g. that all event records link to a patient). Research-quality-level validation is the last level and covers the actual content of the data. CPRD provides a patient-level data quality metric in the form of a binary 'acceptability' flag. This is based on recording and internal consistency of key variables including date of birth, practice registration date and transfer out date.

Separately, a derived death date (consolidating death-related information captured in different parts of the patient record) is currently undergoing validation against both GP-recorded and official ONS death records, and a practice-level quality metric (ascertaining temporal gaps in recording quality) is in development that will be added to future builds of CPRD Aurum.

Weaknesses

Missing data

Though secondary care data, including key diagnoses, can be manually recorded by GPs, this information is often incomplete in primary care records. Additional data may be available in free text entries or letters received by GPs from secondary care facilities, but this is not available to CPRD or researchers for data governance reasons. However, additional information on patient pathways can be obtained through linkage to other data sources as described above. Data on prescriptions for medications and devices that have been issued in primary care are very reliable, however, information on medications dispensed, secondary care prescriptions and over-the-counter use are not recorded in primary care.

GP IT systems and coding

Possible variations in coding between practices and over time, as well as the current transition to SNOMED coding,[21] should be considered by researchers when planning a study using CPRD Aurum. Additionally, the database structure of CPRD Aurum differs somewhat from other UK databases, including CPRD GOLD,[6] due to underlying differences between EMIS Web® and Vision® software structures, which may affect data comparability. CPRD has published preliminary guidance for researchers on differences between the CPRD GOLD and CPRD Aurum databases.[26]

Data resource access

Researchers can apply for a limited licence to access CPRD data for public health research, subject to individual research protocols meeting CPRD data governance requirements. CPRD Aurum data is provided in tab-delimited text files and can be imported into any standard statistical software package. As a not-for-profit organization, CPRD recoups its costs through research user licence fees (annual multi-study license or dataset-specific license), with additional fees for linkage to other datasets.

More details including the data specification, applications process, and access to linked data, are available on the CPRD website (https://www.cprd.com). Researchers can also request feasibility counts from CPRD to inform sample-size estimates and decisions regarding suitability of CPRD Aurum for their proposed research. Any other queries can be directed to CPRD Enquiries [enquiries@cprd.com].

Profile in a nutshell

- CPRD Aurum is a UK primary care database set up for public health research and benefit, updated monthly for observational research, with standard linkages to hospital, mortality, cancer, mental health and deprivation data.
- As of September 2018, 738 GP practices in England have contributed data, which included 19 million patients, of whom over 7 million were currently registered at contributing practices.
- De-identified coded primary care data are collected from GP practices that have consented to provide data to CPRD Aurum.
- Symptoms, diagnoses, prescriptions, immunizations, tests and lifestyle factors are recorded by the GP or other practice staff, and CPRD Aurum has been linked to additional secondary care databases.
- Access to data for public health research is subject to data governance requirements and contractual obligations being met. Queries can be directed to CPRD Enquiries (enquiries@cprd.com).

Conflict of interest: None declared.

References

1. Franks P, Clancy CM, Nutting PA. Gatekeeping revisited—protecting patients from overtreatment. *N Engl J Med* 1992;**327**:424–29.
2. NHS Digital. *General and Personal Medical Services, England: Final 31 December 2017 and Provisional 31 March 2018, Experimental Statistics*. 2018. https://digital.nhs.uk/data-and-information/publications/statistical/general-and-personal-medical-services/final-31-december-2017-and-provisional-31-march-2018-experimental-statistics (16 July 2018, date last accessed).
3. NHS Digital. *Linked Datasets Supporting Health and Care Delivery and Research*. 2018. https://digital.nhs.uk/services/data-access-request-service-dars/linked-datasets-supporting-health-and-care-delivery-and-research (4 September 2018, date last accessed).
4. NHS England. *Notes on GP IT Systems (August 2016)*. 2016. https://www.england.nhs.uk/ourwork/accessibleinfo/resources/gp-it-systems/ (4 September 2018, date last accessed).
5. Kontopantelis E, Stevens RJ, Helms PJ, Edwards D, Doran T, Ashcroft DM. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open* 2018;8:e020738.
6. Herrett E, Gallagher AM, Bhaskaran K *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;**44**:827–36.
7. NHS Digital. *National Data Opt-Out Programme*. 2018. https://digital.nhs.uk/services/national-data-opt-out-programme (25 September 2018, date last accessed).
8. NHS Digital. *[MI] National Data Opt-Out, September 2018*. 2018. https://digital.nhs.uk/data-and-information/publications/statistical/national-data-opt-out/september-2018 (27 September 2018, date last accessed).
9. Ministry of Housing C& LG. *English Indices of Deprivation 2015—GOV.UK*. 2015. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 (19 September 2018, date last accessed).
10. Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *Eur J Epidemiol* 2019; **34**:91–99. doi:10.1007/s10654-018-0442-4.
11. Office for National Statistics. *Deaths Registration Data*. 2018. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths (19 September 2018, date last accessed).
12. Harshfield A, Abel GA, Barclay S, Payne RA. Do GPs accurately record date of death? A UK observational analysis. *BMJ Support Palliat Care* 2018; doi:10.1136/bmjspcare-2018-001514.
13. 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**:1093–1093i.
14. NHS Digital. *Hospital Outpatient Activity*. 2018. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity (19 September 2018, date last accessed).
15. NHS Digital. *Hospital Accident & Emergency Activity*. 2018. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident--emergency-activity (19 September 2018, date last accessed).
16. NHS Digital. *Diagnostic Imaging Data Set*. 2018. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/diagnostic-imaging-data-set (19 September 2018, date last accessed).
17. Public Health England. *National Cancer Registration and Analysis Service (NCRAS)*. 2017. https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras (19 September 2018, date last accessed).
18. Office for National Statistics. *2011 UK Townsend Deprivation Scores*. 2018. https://www.statistics.digitalresources.jisc.ac.uk/dataset/2011-uk-townsend-deprivation-scores (19 September 2018, date last accessed).
19. Wheeler B. *Carstairs Index 2011 for LSOAs*. Colchester, UK: UK Data Service. doi:10.5255/UKDA-SN-851497.
20. Office for National Statistics. *2011 Rural/Urban Classification*. 2013. https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclassifications/2011ruralurbanclassification (19 September 2018, date last accessed).
21. NHS Digital. *SNOMED CT Implementation in Primary Care*. 2018. https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct/snomed-ct-implementation-in-primary-care (19 September 2018, date last accessed).

22. NHS Digital. *Read Codes*. https://digital.nhs.uk/services/terminology-and-classifications/read-codes (19 September 2018, date last accessed).
23. NHS Business Services Authority. *Dictionary of Medicines and Devices (dm+d)* |. https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd (19 September 2018, date last accessed).
24. Clinical Practice Research Datalink. *CPRD Bibliography*. 2018. https://www.cprd.com/Bibliography/ (4 September 2018, date last accessed).
25. Office for National Statistics. *Mid-2017 Estimates of the Population for the UK, England and Wales, Scotland and Northern Ireland*. 2018. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland (19 September 2018, date last accessed).
26. Booth H, Dedman D. *CPRD Aurum Frequently Asked Questions (FAQs)*. 2017. https://www.cprd.com/primary-care (25 September 2018, date last accessed).