Medicines & Healthcare products Regulatory Agency

CPRD

The Public Health England National Cancer Registration and Analysis Service (NCRAS) and CPRD primary care data

Documentation (set 21)

Version 10.1

Date: 26 March 2021

NIHR | National Institute for Health Research

Documentation Control Sheet

Over time, it may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made.

Version	Affected Areas Summary of Change	Prepared By	Reviewed By
1.0	Initial Draft	Rachael Boggon	Arlene Gallagher
1.1	Modified	Rachael Boggon	Helen Strongman
2.0	Modified	Rachael Boggon	Jenny Campbell
2.1	Modified	Rachael Boggon	Kendal Chidwick
2.2	Modified	Rachael Boggon	Kendal Chidwick
2.3	Modified	Rachael Boggon	Dervla Mahoney
2.4	Modified	Grant Lee	Rachael Williams
3.1	Modified	Rachael Williams	Dervla Mahoney
3.1a	Formatted	Grant Lee	Rachael Williams
4.0	Modified	Rachael Williams	Sophia Amjad
5.0	Modified	Rachael Williams	Sophia Amjad
5.1	Modified	Rachael Williams	Sophia Amjad
5.2	Modified	Rachael Williams	Sophia Amjad
6.0	Modified	Rachael Williams	Shivani Padmanabhan
7.0	Modified	Helen Strongman	Rachael Williams
7.1	Modified	Helen Strongman	Eleanor Yelland
8.0	Modified	Rachael Williams	Eleanor Yelland
8.1	Modified	Eleanor Yelland/ Helen Booth/ Tarita Murray-Thomas	Eleanor Yelland/ Helen Booth/ Tarita Murray-Thomas/ Thamina Anjuman
9.0	Modified	Sonia Coton/ Eleanor Yelland	Tarita Murray-Thomas/Eleanor Yelland
10	Modified	Eleanor Yelland	Hilary Shepherd
10.1	Modified	Hilary Shepherd	Susan Hodgson

Version 1.1
- Added ‘Estimating patient follow-up’

Version 2.0
- Updated to match 2010 format

Version 2.1
- Added comments to reg_tmn_X fields and requirement for explicit list of variables

Version 2.2
- Added more details on variable selection form, sign off form, and change of coverage period

Version 2.3
- Added paragraph on data governance to section ‘How do I get access…’
- Renamed type5 to reg_type5 and diag_date to diag_date_format
- Added details on the SLSP
- Replaced references to encryption with pseudonymisation

Version 2.4
- Updated variable names to match dataset agreement spreadsheet

Version 3.1
- Updated for set 10

Version 3.1a
- Formatted with new agency branding

Version 4.0
- Updated for set 11

Version 5.0
- Updated for set 12

Version 5.1
- Updated data structure following further information from PHE

Version 5.2
- Updated stage_path_pretreated following information from PHE

Version 6.0
- Updated document version number, date and set
- Updated coverage period
- Updated references to reflect change of name from HSCIC to NHS Digital

Version 7.0
- Updated document version number, date and set
- Updated coverage period
- Renamed sections to align with other CPRD datasets
- Document renamed, and sections modified to cover three datasets collected through NCRAS (cancer registration data, SACT data, and CPES data)
- Additional advice provided about completeness of cancer registration data

Version 7.1
- Additional advice provided about concordance between cancer registration data and CPRD primary care data, completeness of NCRAS data fields and how to access the NCRAS data.

Version 8.0
- Updated for set 16
- Updated to include CPRD Aurum

Version 8.1
- Updated for set 17
- Additional advice on value of using NCRAS data, on ICD coding
- Addition of QoL PROMS information

Version 9.0
- Updated for set 18

Version 10.0
- Updated for set 19

Version 10.1
- Updated for set 21, removed information on CPES, QOLC and QOLP as these are no longer to be offered as standard linkages
- Updated branding

National Cancer Registration and Analysis Service (NCRAS) data linked to CPRD primary care data

This document provides an overview of the cancer data, as provided by Public Health England (PHE) via the National Cancer Registration and Analysis Service (NCRAS), and the available subset that are linked to CPRD GOLD and CPRD Aurum. The release of cancer data linked to CPRD primary care data (set 21) includes:

- Cancer registration data from January 1990 to December 2018
- Systemic Anti-Cancer Treatment (SACT) for patients with tumours recorded in the cancer registration data from January 2014 to December 2018.
- National Radiotherapy Dataset (RTDS) for patients with tumours recorded in the cancer registration data from April 2009 to December 2018.

How is cancer data collected in England?

Cancer registration in England is supported by a national cancer data standard, the Cancer Outcomes and Services Dataset (COSD). Cancer registrations are centrally managed by the NCRAS within PHE. NCRAS was launched in April 2013, as a single, unified registration service for England. It uses a single national data processing system (English National Cancer Online Registration Environment (ENCORE)) to collect the COSD.

The COSD brings together cancer registration data collected from a range of health care provider systems and other services. These include patient administration systems, multidisciplinary team reporting software, pathology reports, imaging systems, and death certification data. It incorporates elements of the National Cancer Waiting Times Monitoring Dataset (NCWTMDS), items from the Systemic Anti-Cancer Therapy Dataset (SACT) and the National Radiotherapy Dataset (RTDS), all of which also remain as separate standards. All patients diagnosed with or receiving cancer treatment in or funded by the NHS in England are covered by the standard. This includes adult and paediatric cancer patients.

To maximise data quality, historical records are compared and updated as new data become available from different sources. Once all the expected records for any incidence year have been received and validated, the NCRAS take a ‘snapshot’ (a static copy) of the dataset to create the analytical dataset. The analytical dataset is then linked to other administrative datasets on an analytical platform called the Cancer Analysis System (CAS). This linked analytical dataset is available for secondary use purposes and disclosed for research, clinical audit and service evaluation, where appropriate information governance controls are met.

What are the cancer registration data?

The ENCORE data in CAS contain records for all registrable tumours diagnosed or treated in England, of which the NCRAS has been notified. Estimates of registration data ascertainment are very high as the registry is population based and receives death certificates. Cancers are coded using the International Classification of Diseases for Oncology, version 2. They are also back mapped to the tenth revision of the International Classification of Diseases version 10.

Registrable conditions are broadly all invasive tumours, all uncertain behaviour tumours, all in situ tumours and benign tumours within the brain or CNS. Registrable tumours (ICD-10 classifications):

- C00-C97 All malignant neoplasms,
- D00-D09 All in situ neoplasms,
- D32.0 Benign neoplasm of cerebral meninges,
- D33 Benign neoplasm of brain & other CNS,
- D35.2-D35.4 Benign neoplasm of pituitary gland, craniopharyngeal duct and pineal gland,
- D37-D48 Neoplasms of uncertain behaviour,
- E85.9 Primary amyloidosis[1].

The ENCORE data provided by PHE for linkage to CPRD primary care data consist of a single dataset combining information from the CAS for cases recorded from 1990 to 2018. Please note that variables other than the tumour site and date of diagnosis are unlikely to be complete for cases recorded prior to 1995. For simplicity, the dataset provided by PHE for linkage to CPRD primary care data will be referred to as the cancer registration data from this point onwards.

Records for each tumour include fields describing diagnosis and histopathology, operations, procedures and interventions. NCRAS processes all treatment information occurring during the first 6 months after diagnosis. Later treatment data is collected and processed but may not always be available. (Henson et al., 2019)

What is the Systemic Anti-Cancer Therapy Dataset (SACT)?

The SACT dataset covers systemic anti-cancer treatment for all solid tumour and haematological malignancies, including those in clinical trials. Treatments include hormones and bisphosphonates, oral chemotherapy, BCG/intravesical chemotherapy, and targeted/biological therapies. Information is included about programme and regime of treatment, and the outcome for each treatment. The dataset includes all cancer patients, both adult and paediatric, treated in acute inpatient, day case outpatient and community settings funded by the NHS in England.

Phased implementation of SACT began in 2012 and submission by all trusts in England became mandatory in April 2014. By January 2014, 95% of trusts were routinely submitting at least the mandatory data items (Wallington et al., 2016). SACT ascertainment can therefore be considered largely complete for patients with tumours recorded in the cancer registration data from January 2014 to December 2018. Treatments administered for these tumours are available approximately 6 months behind the current date. SACT data can also be requested for patients with tumours recorded in the cancer registration data between April 2012 and December 2013, noting that ascertainment was incomplete during this period. Please note that SACT data is linked to cancer registration data at the patient level, but not the tumour level.

Research teams using SACT data should be aware that PHE is unable to release SACT data for patients who are receiving a subset of treatments currently funded by the Cancer Drugs Fund. Records for such patients will be absent from all NCRAS data (including cancer registration and other NCRAS datasets in addition to SACT) for a study-specific dataset, if SACT data is requested. Please contact the CPRD Observational Research Team on enquiries@cprd.com if you would like to discuss how this may impact your research.

[1] Primary amyloidosis is widely recognised and treated as a cancer by clinicians. Its inclusion as a registerable condition has been agreed with NCRAS. It is expected that the WHO disease classification will be amended to reflect this in time.

Treatments are organised into programmes, regimens and cycles. Drug programmes equate to a line of chemotherapy and are numbered chronologically. Regimens are groups of drugs administered in a specific way according to various timings and parameters. Mostly the programmes and regimens will be the same, but for paediatric and haematological cancers multiple regimens may be given together or sequentially and comprise one programme. Cycles are repeating patterns of drugs being administered to patients within a regimen.

What is the National Radiotherapy Dataset (RTDS)?
The RTDS dataset contains records of radiotherapy services provided since April 2009, including teletherapy and brachytherapy. All radiotherapy delivered in England to patients in NHS facilities, or in private facilities where delivery was funded by the NHS, is included. Brachytherapy delivered for the treatment of non-malignant disease, radiotherapy delivered using unsealed sources, and non-therapeutic exposures delivered using radiotherapy machines (e.g. imaging) are not included.

Treatments are organised into episodes, with each episode representing a continuous period of care for radiotherapy including all preparation, planning and delivery of treatment as covered in the treatment intention. Treatment given concurrently or consecutively to multiple sites associated with the same primary tumour will form part of the same episode. If the treatment plan changes during treatment, then all of the treatment delivered concurrently/consecutively will form part of the same episode. Treatment given to separate unrelated primary tumour sites will form separate episodes.

Other cancer datasets
Cancer Patient Experience Survey (CPES)
The National Cancer Patient Experience Survey (CPES), commissioned by NHS England through Quality Health, is a survey sent out to all adult cancer patients (aged 16 and over) with a primary diagnosis of cancer who have been admitted to an acute or specialist NHS Trust in England providing adult cancer services as inpatients or day cases, and discharged within a specified three-month sampling period each year. The survey aims to collect information from patients about their cancer journey from their initial GP visit prior to diagnosis, through diagnosis and treatment and to the ongoing management of their cancer. The survey is conducted in waves each, by recruiting patients who were discharged over the course of three months in a given year. Nine waves are currently available for linkage covering years 2010 to 2019. The specific questions asked can vary from year to year (see data dictionary).

The CPES dataset is no longer routinely available for linkage to CPRD data but may be made available for specific projects. Please contact the CPRD Observational Research Team on enquiries@cprd.com to discuss your requirements.

Quality of Life of Cancer Survivors in England: Pilot Patient Reported Outcome Measures Survey (2011) (Breast, Colorectal, Prostate, Non-Hodgkin's Lymphoma)

The Quality of Life of Cancer Survivors in England: Pilot Survey (2011) was commissioned by the Department of Health as part of the National Cancer Survivorship Initiative (NCSI). The survey was conducted by Quality Health in conjunction with three cancer registries in England.
The aims of the survey were to assess the feasibility and acceptability to cancer survivors of collecting information on quality of life (QoL) using Patient Reported Outcome Measures (PROMS), to assess the overall quality of life of representative samples of cancer survivors with four different incident tumour types at four different time points after diagnosis, and to assess the contribution of demographic, disease-related, and other, factors to quality of life.
The survey measured the overall quality of life of representative samples of cancer survivors with breast, colorectal cancer, prostate cancer and non-Hodgkin's lymphoma (NHL) diagnosed during July 2006 - July 2010. Quality of life was assessed at four different time points after diagnosis at approximately one, two, three or five years.
An initial sample of 4,992 patients was identified by three participating cancer registries (West Midlands, East of England and Thames). This sample comprised 312 patients in each of the four tumour groups at each of the four time points. The overall response rate was 66%. Response rates were broadly similar across the tumour groups and between time points, with around 200 patients responding in each tumour group at, each time point.

Outcome questions in the survey are made up of three instruments: the EQ-5D (Euroqol 5 level), FACT items (Functional Assessment of Cancer Therapy) and SDI (Social Difficulties Inventory).
Other survey items used for all tumour groups were selected to cover demographic information, treatment details, disease status (e.g. remission, relapse, uncertain), long term conditions (one item), physical activity (one item), social difficulties (11 items), psychological issues (nine items), work status (one item) and experience of care (six items). A total of 43 questions were common to all tumour groups, with around 20-30 additional questions depending on tumour type (see data dictionary).

As these data describe the quality of survival of a sample of people with four different cancers, they provide useful insights on the consequences of survival and impact on function, factors that impact on outcome, including treatment and may inform appropriate health and social care, clinical trials and supportive care research.

Further information on the findings of this survey can be found at
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/267042/9284-TSO-2900701-PROMS-1.pdf

The QOL dataset is no longer routinely available for linkage to CPRD data but may be made available for specific projects. Please contact the CPRD Observational Research Team on enquiries@cprd.com to discuss your requirements.

Quality of Life of Colorectal Cancer Survivors in England: Patient Reported Outcome Measures Survey (PROMS)

The Quality of Life of Colorectal Cancer Survivors in England: Patient Reported Outcome Measures survey, was commissioned by the Department of Health as a follow-on from the pilot study in July 2011 undertaken to confirm the value of collecting PROMs data on breast, prostate, colorectal and non-Hodgkin's lymphoma. (See the section above on "*Quality of Life of Cancer Survivors in England: Pilot Patient Reported Outcome Measures Survey (2011) (Breast, Colorectal, Prostate, Non-Hodgkin's Lymphoma)*)".

It is a **national** survey conducted among patients aged 16 years and over with an incident colorectal cancer diagnosis during Jan 2010 - Dec 2011. Colorectal cancer was defined as cancer of the colon, cancer of the rectosigmoid junction or cancer of the rectum, as per the International Classification of Diseases register (ICD-10 codes C18-C20). Only persons receiving treatment in the National Health Service in England were included.

In January 2013, 34,467 patients who were alive 12-36 months after diagnosis were asked to complete a questionnaire about different parts of their patient journey from their colorectal cancer diagnosis to treatment through to their experiences of aftercare. The questionnaire comprised of 76 questions (see data dictionary), with an additional comments box (data not available from the CPRD) and were divided into several different sections as outlined below. Outcome questions in the survey were made up of three instruments: the EQ5D (Euroqol 5 level), FACT items (Functional Assessment of Cancer Therapy) and SDI (Social Difficulties Inventory). The response rate in this questionnaire survey was 63%.
As these data describe the quality of survival of people with colorectal cancer, they can be useful in identifying consequences of survival and impact on function, identifying factors that impact on outcome, including treatment; comparing outcomes by service provider organisations, supporting enhanced delivery of care, enabling provision of appropriate health and social care and clinical trials and supportive care research.

Sections of the Quality of Life of Colorectal Cancer Survivors in England: Patient Reported Outcome Measures Survey (PROMS)

- **General questions** - a range of questions asking about the type of treatment the person had, the length of time since their treatment, how well their cancer had responded to treatment and whether they had a stoma or not.
- **Outcome questions** - a set of questions using three different instruments to assess how the patient felt about the impact of the cancer physically and emotionally (Q5-9 for EQ5D 5L questions, Q10-22 for FACT items and Q26-46 for SDI items
- **Overall support and care** - several questions about the care the person received in primary care (GP/community care) and secondary care (hospital), as well as questions around access to information and support, and more general questions about lifestyle (smoking and exercise).
- **About you** - demographic questions were included to enable the results to be considered alongside factors such as age, sex, deprivation, ethnicity, and presence of long-term conditions.
- **Comments** - a free text box was available for patients to make any additional comments on any aspects of living with cancer not touched on elsewhere in the survey or to provide further views and explanations. **Please note that these data are not available in CPRD**

Further information on the findings of this survey can be found at the links below:

https://www.england.nhs.uk/wp-content/uploads/2015/03/colorectal-cancer-proms-report-140314.pdf

https://eprints.soton.ac.uk/374712/1/Health%2520Related%2520Quality%2520of%2520Life%2520After%2520Colorectal%2520Cancer.pdf

The QOL PROMS dataset is no longer routinely available for linkage to CPRD data but may be made available for specific projects. Please contact the CPRD Observational Research Team on enquiries@cprd.com to discuss your requirements.

Linkage algorithm and the match rank variable

Patients in CPRD primary care data are linked to the NCRAS data using an eight-step deterministic algorithm based on four identifiers, shown in Table 1 below. The linkage is undertaken by NHS Digital, acting as a trusted-third-party, on behalf of CPRD. No personal identifiers are held by CPRD, or included in the CPRD GOLD, CPRD Aurum, or linked NCRAS data.

Table 1: NHS Digital 8 step linkage algorithm

Step	Match	CPRD GOLD Percent	CPRD Aurum Percent
1	Exact NHS number, sex, date of birth (DOB), postcode	57.4	52.8
2	Exact NHS number, sex, DOB	37.4	41.5
3	Exact NHS number, sex, postcode, partial DOB	0.7	0.7
4	Exact NHS number, sex, partial DOB	0.6	0.6
5	Exact NHS number, postcode	0.1	0.1
6	Exact sex, DOB and postcode (where NHS number does not contradict the match, the DOB is not 1st of January & the postcode not on the communal establishment list)	3.6	4.0
7	Exact sex, DOB and postcode (where the NHS number does not contradict the match and the DOB is not 1st of January)	0.2	0.2
8	Exact NHS number	0.1	0.1

The matching steps are applied sequentially. If a CPRD GOLD or CPRD Aurum patient record is matched in one step, it is no longer available for matching in subsequent steps. The table above shows the percentage of records that were matched at each step for the cancer registration data in set 21. The SACT, RTDS, CPES, and QoL PROMS data are linked to the cancer registration data by PHE.

CPRD provides users with a match_rank variable which corresponds to the step at which the match was established. In general, a lower value for the match_rank is considered stronger evidence for a positive match. Note that only patients with a match_rank of 5 or less are considered definitive matches and are included in standard linked NCRAS datasets. Please speak to a member of the CPRD Observational Research team prior to submission of your study protocol if you would like to access data matched in ranks 6 to 8.

Not all patients in CPRD primary care data are eligible to be linked to NCRAS data, for example, due to the region in which they reside (outside England), the lack of a valid NHS identifier, their GP practice not having consented to linkage, or if they have personally dissented from their records being used for linkage purposes. A study specific population source file and accompanying documentation will be provided to researchers with the dataset. This will enable identification of the subset of those CPRD primary care patients in the study who were eligible to have a record in the cancer registration data during the available coverage period.

Known issues and guidance on the use of linked cancer registration data

To the maximum extent possible, CPRD has left the entries in the fields of the cancer registration data "as is". As such, there are a number of data issues that researchers should be aware of, described below.

(i) Multiple links between identifiers

One cancer registration tumour identifier can be linked to many CPRD patient identifiers. This can occur when, for example, a patient moved from one CPRD practice to another, which CPRD are unable to identify. Researchers will receive each CPRD primary care patient in their study that a cancer registration patient has been linked to, and potential duplicate patients will need to be handled on a study by study basis. It is also possible that multiple cancer registration patient identifiers can be linked to one CPRD patient identifier, for example if two registry entries were made for the same person at different times and a new cancer registry ID was generated for the same patient.

(ii) Tumour level data

Cancer registration data are collected and recorded at the individual tumour level. This means that patient level variables (e.g. sex) are recorded multiple times and can vary between different tumour records for the same patient. As such, CPRD recommends using sex as recorded in the CPRD primary care data, rather than that recorded in the cancer registration data.

SACT data are linked to cancer registration data at the patient level. Different tumour identifiers are used in the two datasets.

(iii) Concordance between cancer registration and primary care data

Concordance of recording on incident tumours in cancer registration data and primary care data has been shown to be high for the majority of tumour types. Diagnosis dates in the cancer registration data are commonly a number of days earlier than the first record in the primary care data (Boggon et al., 2013; Dregan et al., 2012). Cancer registration data also provides better differentiation between primary tumours and metastases than the primary care data. Feasibility counts for study protocols can either be based on the primary care data alone, published manuscripts, or requested from PHE through the CPRD Observational Research Team. Feasibility counts from PHE are provided as a range indicating the approximate number of patients with a tumour record for a specified ICD-10 code list over a specified calendar period of time. It is not possible to restrict these to tumours occurring during continuous follow-up in the CPRD practice.

(iv) Ascertainment of treatment records in the SACT data

When compared to the Cancer Waiting Times (CWT) dataset (which also contains chemotherapy as a selection of treatment categories along with treatment start date) the SACT dataset was found to have a high level of completeness with regards to patient treatment (around 88% of patients reported to have received chemotherapy during 2014 were included in the dataset) (National Cancer Registration and Analysis Service, 2014).

However, endocrine treatments (which are key treatment modalities for many patients with breast or prostate cancer) were found to be significantly under-reported for patients. More than 80% of these treatments that had been reported through CWT were missing in the SACT dataset. This may be because they are commonly prescribed in primary care. As mentioned previously, coverage of hospital trusts is incomplete prior to 2014.

In addition, records of relatively new treatments funded by the Cancer Drugs Fund at the time of the data request will not be available to researchers. The impact of this limitation upon the representativeness of your study should be considered and discussed with the CPRD Observational Research Team prior to submitting a protocol through the RDG (research data governance) process on eRAP (electronic research applications portal).

(v) Level of response to CPES

A comparison of respondents from the CPES survey and the cancer registration data in CAS indicated a high concordance of basic patient characteristics in the two datasets. Please refer to the

report for further information about patterns of variation (National Cancer Intelligence Network, 2015). Patients are invited to participate within specific time windows and non-response has been highlighted as an issue in a published study (Pinder et al., 2016).

(vi) Completeness of data fields in the NCRAS data

Completeness of data fields in the cancer registration data varies significantly by tumour type and over time. Information on cancer registration data completeness is available on CAS Explorer website: https://www.cancerdata.nhs.uk. Advice and feasibility counts can be sought from PHE through CPRD when the level of completeness of individual fields will affect study feasibility.

(vii) How can I find out more?

Further information about the COSD can be found on the NCRAS website (http://www.ncin.org.uk/home). PHE's e-learning hub also provides useful modules describing the evolution of cancer data collection in the UK (http://www.mylearningspace.me.uk/moodle/). A synthetic dataset, imitating elements of the cancer registration and SACT data, has been produced by Health Data Insight CiC. The Simulacrum may be a useful resource for learning about the NCRAS data (see details in the section **'Is NCRAS data suitable for my study?'** below).

Look-up files

CPRD will not be providing ICD or OPCS dictionaries for use with linked cancer registration data. The ICD codes have been slightly modified from those provided by the World Health Organisation (WHO). The cancer registration data use the ICD 10-O2 coding and requests should be based on the four-digit site code. CPRD recommend acquiring lookup tables for ICD codes from the NHS Digital Clinical Classifications Service by emailing them at information.standards@nhs.net or by telephoning 0845 13 00 114. Note that a license is required.

It is likely that the lookup table that will be of most use to you is the ICD Metadata file. This file contains all valid ICD codes, their titles, and category titles together with age and gender validation flags. You will be able to find out further information, including details of the license you will need to obtain at: https://digital.nhs.uk/article/1117/Clinical-Classifications

The Office of Population Census and Surveys (OPCS) Classification of Procedures and Interventions codes are also available from the NHS Digital Clinical Classifications Service. As with ICD codes, a license may be required.

Data dictionaries for SACT and RTDS data are also available on the NHS Data Dictionary website: https://www.datadictionary.nhs.uk/

Is NCRAS data suitable for my study?

The data held by NCRAS provide more detail on the tumour diagnosis, treatment and experience of cancer patients than is available in primary or linked HES data. Recent work has suggested that for case ascertainment, cancer diagnoses should ideally be based on information in the NCRAS cancer registration data. It was also suggested that patient demographics and the route of diagnosis of cancer are likely to impact the accuracy of cancer recording in primary care data (Arhi et al., 2018). Data resource profiles for the NCRAS Cancer Registry (Henson et al., 2019) and SACT (Bright et al., 2020) data were recently published and will be valuable in understanding the data available. Researchers should be aware that NCRAS data is not held at CPRD and therefore needs to be requested on a study by study basis. It is also subject to additional data governance requirements. These factors lead to longer processing and delivery times, and access to NCRAS data is subject to an additional one-off cost.

An additional resource that may be helpful in determining whether NCRAS data would be of value to your research is The Simulacrum. The Simulacrum was developed by Health Data Insight (HDI) CiC in partnership with AstraZeneca (AZ) and IQVIA. It contains artificial patient-like cancer data that imitates some of the data held by NCRAS. It is free to use and does not contain any information about real patients. For further information about the Simulacrum and how to access it please see: https://healthdatainsight.org.uk/project/the-simulacrum/

How do I get access to the cancer registration data?

PHE are the custodians of cancer data held in the CAS. Requests for CPRD primary care linked NCRAS data can only be made via the CPRD. Data for such projects will be provided as study-specific datasets in which all patient, staff and practice identifiers are re-pseudonymised with identifiers specific to the study.

The cancer registration (CR), SACT, RTDS, CPES, and QoL PROMS data can be accessed, on a study by study basis, as an additional set of data linked to CPRD primary care data. Data are made available to the CPRD by the PHE in separate files. These files can then be linked to the corresponding CPRD primary care data using the pseudonymised CPRD patient identifier (e_patid). There is a fixed cost for receiving linked NCRAS data. This is levied to offset the costs that the CPRD and PHE incur from setting up and enabling the linkage and providing related services.

Upon deciding that linked NCRAS data are required for a study, a protocol indicating the request should be drafted and discussed with CPRD's Observational Research team. Please indicate which cancer datasets you require in the application: CR data with or without SACT, RTDS, CPES and / or QoL PROMS data. The protocol should explain why linked cancer data are required and what these will be used for.

The NCRAS Data Selection Form (available from CPRD as a separate Excel spreadsheet) should be completed by selecting the variables that are required to meet the objectives of the study as described in the study protocol. The study protocol will need to contain details of how each selected variable is to be used in the proposed study. CPRD will be unable to request or release variables that have not been justified in the protocol. The CPRD Observational Research Team will review the application ensuring that you have demonstrated that the variables are necessary to meet the study objectives and that you have demonstrated this in the methodology sections of the protocol. Applications containing requests for variables that are not explicitly justified in the protocol will be returned to applicant and require further review. Once the Observational Research team member has confirmed by email that these requirements have been met, the applicant will be provided with a reference number and finalised version of the NCRAS DSF. The protocol can then be submitted on eRAP (electronic research applications portal) with this information.

Data shall only be processed in the following Territory:

- UK
- EEA
- Non-EEA–restricted to Argentina, Australia, Canada, Faroe Islands, Guernsey, Isle of Man, Israel, Jersey, New Zealand, Switzerland, United States, Uruguay

No patient or tumour level data provided under this Contract will be processed outside the European Economic Area without the prior approval of the PHE Office for Data Release. Any and all such processing activities, including but not limited to, international transfers will be reviewed on a case by case basis by the PHE.

Requesting additional approval from PHE to use the NCRAS data outside of these areas may impact on timelines for access to the data.

Once approval has been obtained through the CPRD RDG (research data governance) process, and PHE if necessary, CPRD will work with the client in order to define the data requirements. CPRD will manage the data extraction and delivery process and liaise with PHE on behalf of the client. Please note that the approved protocol and any subsequent amendments will be shared with PHE. Patient, practice and staff identifiers in the primary care data, and all other linked data, will be replaced by pseudonymised identifiers generated by CPRD.

Due to data governance issues, clients must not prepare a patient list and send this to CPRD requesting cancer registration data. Cancer registration data can only be provided when CPRD have

produced the patient list on behalf of the client and pseudonymised the data. This process, along with standard clauses in the CPRD license documentation, aims to prevent clients from matching Cancer Registration data to CPRD data in its original format. Clients with access to CPRD data for other studies, either via online access or study specific datasets, should not attempt to merge this data with cancer registration data, as this would contravene the license agreement.

Future plans

Additional calendar years of cancer registration data will be incorporated as they become available, as will additional cancer related data from the CAS. Plans have been made to include data from Wales, Scotland and Northern Ireland, but there is no timescale set on when this might happen. The cancer registration to CPRD primary care link is part of the total linkage programme that will enable, over time, highly detailed and complete, anonymised, longitudinal patient journeys to be tracked.

References

Arhi, C. S., Bottle, A., Burns, E. M., Clarke, J. M., Aylin, P., Ziprin, P., & Darzi, A. (2018). Comparison of cancer diagnosis recording between the Clinical Practice Research Datalink, Cancer Registry and Hospital Episodes Statistics. *Cancer Epidemiology*, *57*(May), 148–157. https://doi.org/10.1016/j.canep.2018.08.009

Boggon, R., van Staa, T. P., Chapman, M., Gallagher, A. M., Hammad, T. A., & Richards, M. A. (2013). Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiology and Drug Safety*, *22*(2), 168–175. https://doi.org/10.1002/pds.3374

Bright, C. J., Lawton, S., Benson, S., Bomb, M., Dodwell, D., Henson, K. E., McPhail, S., Miller, L., Rashbass, J., Turnbull, A., & Smittenaar, R. (2020). Data Resource Profile: The Systemic Anti-Cancer Therapy (SACT) dataset. *International Journal of Epidemiology*, *49*(1), 15–15l. https://doi.org/10.1093/ije/dyz137

Dregan, A., Moller, H., Murray-Thomas, T., & Gulliford, M. C. (2012). Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer Epidemiology*, *36*(5), 425–429. https://doi.org/10.1016/j.canep.2012.05.013

Henson, K. E., Elliss-Brookes, L., Coupland, V. H., Payne, E., Vernon, S., Rous, B., & Rashbass, J. (2019). Data Resource Profile: National Cancer Registration Dataset in England. *International Journal of Epidemiology*. https://doi.org/10.1093/ije/dyz076

National Cancer Intelligence Network. (2015). *English National Cancer Patient Experience Surveys linked to cancer registration data. A descriptive overview of respondents' characteristics*.

National Cancer Registration and Analysis Service. (2014). *Completeness of the national Systemic Anti-Cancer Therapy data set compared with the Cancer Waiting Times data set*.

Pinder, R. J., Ferguson, J., & Møller, H. (2016). Minority ethnicity patient satisfaction and experience: Results of the National Cancer Patient Experience Survey in England. *BMJ Open*, *6*(6), e011938. https://doi.org/10.1136/bmjopen-2016-011938

Wallington, M., Saxon, E. B., Bomb, M., Smittenaar, R., Wickenden, M., McPhail, S., Rashbass, J., Chao, D., Dewar, J., Talbot, D., Peake, M., Perren, T., Wilson, C., & Dodwell, D. (2016). 30-

day mortality after systemic anticancer treatment for breast and lung cancer in England: A population-based, observational study. *The Lancet. Oncology*, *17*(9), 1203–1216. https://doi.org/10.1016/S1470-2045(16)30383-7