Public Health England

Protecting and improving the nation's health

A guide to completing the NCRAS data dictionary

About Public Health England

Public Health England exists to protect and improve the nation's health and wellbeing, and reduce health inequalities. We do this through world-leading science, research, knowledge and intelligence, advocacy, partnerships and the delivery of specialist public health services. We are an executive agency of the Department of Health and Social Care, and a distinct delivery organisation with operational autonomy. We provide government, local government, the NHS, Parliament, industry and the public with evidence-based professional, scientific and delivery expertise and support.

Public Health England
Wellington House
133-155 Waterloo Road
London SE1 8UG
Tel: 020 7654 8000
www.gov.uk/phe
Twitter: @PHE_uk
Facebook: www.facebook.com/PublicHealthEngland

Published November 2020
PHE publications
gateway number: GW-1753

PHE supports the UN
Sustainable Development Goals

Corporate member of Plain English Campaign
Committed to clearer communication
339

Contents

Background

The National Cancer Registration and Analysis Service (NCRAS) holds many linked datasets on cancer patients. This data is made available through the Office for Data Release (ODR) for external projects. The cancer analysis team at NCRAS supports this process by providing pre-application advice to applicants. This helps ensure the data specification meets the applicant's analytical needs, and that the applicant has a full understanding of the complexities of the data and its limitations before applying. The analytical team also reviews cancer ODR applications before they are approved, and is responsible for extracting, quality assuring, and delivering the data according to the approved data specification.

Pre-application – understanding NCRAS data

The cancer analysis team recommends applicants read the following guides before applying to fully understand the cancer data held by PHE:

1) **A guide to NCRAS data and its availability**: This guide aims to explain the flow of data into NCRAS and the differing availability of data items based on their source.
    a. This includes a **lung cancer data availability** appendix outlining how to request the most comprehensive data on lung cancer cases utilising both the cancer registry (from COSD), and separately collected audit data (known as LUCADA from 2005 to 2014 and NLCA from 2015 onwards).
2) **Guide to using the Simulacrum and submitting code:** Guidance outlining how researchers can use the Simulacrum to plan and test their hypotheses before making a formal request to PHE to analyse real patient data.
3) Published data profiles providing detailed insight into specific datasets:
    a. National Cancer Registration Dataset data profile
    b. SACT data profile
    c. HES data profile
4) **Data Standards** for each of the datasets – for further information regarding individual data items

NCRAS data dictionary

The NCRAS data dictionary provides information on the cancer registration data that is available to request and contains field lists and associated metadata. It has specifically been designed to enable users to easily discover what can be requested and to co-produce a final data specification for a specific project, ahead of submission to the ODR.

Please consider the following aspects when completing the NCRAS data dictionary:

Cohort definition

Applicants should carefully consider how to define the cohort of data they require for their study and should only request variables that are essential. In the first instance, they should work through the Mandatory Defining Cohort tab in the Data dictionary which will allow the applicant to select the datasets they require for their analysis.

The applicant is required to provide as much detail as possible as to how the cohort is defined. Where relevant, the applicant should provide the codes required for their cohort ie which topography or morphology codes should be used to define the cancer site of interest, which treatment code(s) should be used to define the treatment of interest, the trust code if applicable or which geographical boundary code they are interested in. Codes may be specified in ICDO3 or ICD10 (please indicate which coding system the specification is using).

Where a cohort covers multiple tumour types or multiple years, applicants should consider whether information on prior or subsequent primaries to the main cancer of interest are needed, for example, do they want all tumours or simply the first (or last) tumour in a set time period or only the first tumour of a particular cancer site.

If requesting information regarding treatment, the applicant should advise whether they are happy to stick with the default time frame around diagnosis as outlined in the data dictionary or if they require treatments for a longer or shorter period of time.

At all stages, applicants should provide as much detail as they can. It is strongly advised that the applicant discuss their requirements with a pre-application analyst to ensure the appropriate level of detail is provided to process the request in a way that is appropriate to the research question(s).

Temporality of the data available

Each dataset contains data for a specific date range, based on the date of diagnosis of a tumour (National Cancer Registration Dataset, NCDA), or the date of a particular event (SACT, RTDS, HES, DID, CPES, CWT).

For example, linked SACT data is available for patients diagnosed prior to the start of SACT data collection (diagnoses from 1 Jan 1995 onwards) as well as after the start of SACT data collection, if those patients have received SACT since 1 April 2012 and been included in the SACT dataset. However, please note linked data for patients diagnosed prior to the start of SACT is most likely to be available for cancer sites with good prognosis.

Full details of the temporality of the data available for request are given in the NCRAS data dictionary.

Patient vs tumour linkage

Data on each event, treatment or outcome in linked tables is joined to information in the National Cancer Registration Dataset at either a patient-level or at the tumour-level (as each patient may have more than one tumour during their lifetime).

Data linked at a patient-level are typically linked using the NHS number of the patient. Algorithms to identify the tumour most likely to relate to each event, treatment or outcome (where appropriate) further build on this linkage by typically utilising information on the date of diagnosis, referral or treatment, as well as the cancer site of the tumour.

Further details regarding whether each dataset is available for request either patient-linked or tumour-linked to the National Cancer Registration Dataset are given in **A guide to NCRAS data and its availability.**

Minimising the amount of data requested

Only the minimal number of data items needed to undertake a specific project (data minimisation) should be requested. Careful thought should be given to whether full dates are needed, or whether pre-calculated intervals or (month and) year would suffice to answer the research question(s).

Fields with restrictions (eg to specific cancer sites)

Some variables are only collected or only routinely completed for a certain type of cancer, for example, the oestrogen receptor score of the tumour is only available for invasive breast cancer (ICD10 C50x) and DCIS (ICD10 D05x), whereas the Gleason score is only applicable to prostate cancer (ICD10 C61x). Information on any such restrictions or applicability is given in the “Key information” column for each variable within the NCRAS data dictionary.

Other fields may only be available for specific time periods due to changes in how the data items are collected over time or when fields became mandatory or available to collect. Some fields may only be available to applicants with specific types of permissions (such as permission to receive identifiable data items).

NCRAS derived fields
NCRAS extracts and converts raw, unstructured data, which contains duplicated values and/or inconsistent values from multiple trust returns, into a single value per time point of interest, through both the process of cancer registration and by developing, testing and finalising specific algorithms.

For example, STAGE_BEST is an NCRAS derived field using a combination of best T, N and M values and where appropriate site-specific staging systems. For most applicants requesting the stage of the tumour, this field is therefore likely to be the best option. However, depending on the analytical question to be answered, the applicant may wish to request alternative stage fields such as the stage at diagnosis derived from imaging (STAGE_IMG) or the pathological stage at diagnosis (STAGE_PATH).

Another example of derived fields is the Route to Diagnosis. Each tumour is assigned a likely route in which it was diagnosed by combining information from various sources including the cancer registration data, Hospital Episode Statistics, Cancer Waiting Times and screening data. Similarly, a comorbidity score for each tumour is calculated by summarising the diagnostic codes in the admitted patient care hospital episode statistics records into the Charlson comorbidity groups.

Differing completeness of fields
The National Cancer Registration Dataset holds information on demographics, tumour staging, treatments and hospital use for each cancer patient. However, different fields have differing levels of completeness, dependent on whether individual providers are submitting data to NCRAS.

The CAS Explorer shows the overall completeness of these fields for all patients, and is designed to be used by researchers and analysts to initially check the data quality of fields before requesting data through ODR.

Researchers can also use the Simulacrum dataset, a synthetic dataset which mimics some of the cancer registration data, and which can be used to undertake a more detailed assessment of the pattern of completeness and distribution for data items for specific tumour and treatment types.

Defining events of interest/how to link events to specific tumours
Different data fields may relate to either a specific patient, a specific tumour, or a specific event. For example, demographic information such as the date of birth, sex and ethnicity are patient-level, whereas information such as the stage at diagnosis relates to a specific tumour, and information regarding a specific treatment or hospital admission relate to specific events.

This data is generally structured within one-to-many relationships between patient, tumours and events. However, not all event-level information has been tumour-linked and is only available linked to an individual patient.

Where necessary, applicants should consider how to define inclusion criteria for event level data which has not been tumour-linked by NCRAS (further detail on the “Defining cohort” tab within the NCRAS data dictionary).

For example, where an applicant wishes to request treatment and hospital admission information on a cohort of patients with breast cancer, it would be necessary to specify how to determine which HES records are requested, for example by requesting all episodes of care within 30 days before to 6 months after diagnosis for patients in the initial cohort defined and if particular types of event are of interest (eg mastectomy, breast conserving surgery, sentinel lymph node biopsy, axilla block dissection) or every event is of interest. Surgical events should be defined by OPCS4 codes.

Differing HES data flows

NCRAS receive two different flows of HES inpatient data depending on the date of diagnosis of the patient - this affects the availability of linked HES inpatient data as follows:

1. Patients diagnosed before 2014: data for individual events is available from 1/4/1997 to the most recent time point available (using HES2015 and HESLIVE).
2. Patients diagnosed after 2013: data for individual events is available from 1/4/2000 to the most recent time point available (using HESLIVE only).
3. Patients who are diagnosed with different tumours both before 2014 and after 2013: data for individual events is available from 1/4/1997 to the most recent time point available.

Please note: For those patients in #1 and #3 who survive a long time with multiple events, the period of overlap between the two HES flows (1/4/2000 to 31/12/2014) will have duplicate events - NCRAS can de-duplicate this data, however this process can be complex depending on the specific algorithm used, which is likely to affect the cost estimate for delivery of the data.

Differing RTDS data flows

The Radiotherapy Data Set (RTDS) standard (SCCI0111) is an existing standard that has required all NHS Acute Trust providers of radiotherapy to return data for all radiotherapy activity. All teletherapy and brachytherapy treatments are submitted that are delivered in England to patients in NHS facilities, or in private facilities where delivery is funded by the NHS, from 1 April 2009.

The National Clinical Analysis and Specialised Applications Team (NATCANSAT) based at The Clatterbridge Cancer Centre NHS Foundation Trust was responsible for the management and delivery of the RTDS standard from April 2009 until the end of March 2016.

Public Health England (PHE) took over full responsibility for RTDS with effect from 1 April 2016. As such, PHE do not have access to the original radiotherapy provider submissions prior to 1 April 2016 and may not be able to answer queries regarding the data quality of these submissions.

Dose and fractionation are captured in two different ways in the data; prescribed and actual dose, and prescribed and actual number of fractions. Historical data completeness of prescribed dose and fractions is more complete than for actual dose and fractions, however, which data fields are most appropriate to request also depends upon what the data will be used to assess.

It is advised that applicants discuss the research or study with the RTDS team through ODR if planning to use the dose and fractionation fields.

Differing SACT data flows

The SACT dataset was established in 2012 (mandated from April 2014) and up until June 2018 it was maintained as a standalone dataset by NCRAS (internally known as the SACT Legacy dataset).

Since June 2018 trusts have been submitting their monthly SACT data submissions to a new portal (internally known as ENCORE SACT). The way SACT data is handled and processed has been improved, enabling better checks on the quality and completeness of data being uploaded through the SACT portal.

Additionally, the new portal brings SACT data into the cancer registration process and enables tumour level linkage to be carried out by NCRAS registration officers.

Cancer Drugs Fund (CDF) restrictions

Since July 2016, PHE have been evaluating new drugs that have entered the CDF to support the NICE committee decision making process. As a result, PHE may temporarily restrict releasing SACT data relating to CDF evaluations if the request is to calculate ‘outcomes’, particularly treatment duration or survival. The restrictions will also depend on the size of the CDF cohort that exists within the SACT data extract for the ODR request. Information regarding cohort characteristics can be released for SACT data relating to CDF evaluations. Each application will be considered on a case by case basis after a risk assessment has been carried out.

These restrictions apply until NICE make a decision as to whether the drug will be routinely commissioned or de-commissioned.

SACT tumour linkage

Tumour level linkage using ENCORE SACT (June 2018 onwards) is tumour linked using an algorithm and manual matching, manual matching is based on the tumour a registration officer has determined is linked to the SACT regimen.

In contrast, for data submitted prior to June 2018, SACT tumour level linkage can only be carried out using a SACT tumour level linkage algorithm developed by the NCRAS analytical team based on the SACT Legacy dataset.

There are known limitations to both mechanisms for SACT tumour level linkage including:

SACT Legacy tumour level linkage limitations:

- the treatment date rules used by the algorithm need to be adjusted for certain diagnostic groups – for example, systemic therapy after one year may be common for one type of cancer but not for another, therefore if the correct date rule is not applied then SACT records will be missing
- patients with very long gaps (more than 2 years) between treatments, may not be picked up, as it is difficult to determine without checking each set of data whether the events are late treatment for the tumour in question or early treatment for a later diagnosis
- where a patient has only one tumour registered eg C50, but the SACT record has a different type of tumour recorded eg C43, the algorithm may not match the records. The primary diagnosis in SACT may have been incorrectly recorded. Please read the SACT data profile
- the tumour level linkage algorithm maps cancer registrations diagnosed from 2010 onwards only. This restriction will therefore miss treatments for pre-2010 tumours with long-term follow up. This will be an issue for diagnoses with long periods of treatment and survival, such as breast cancer

ENCORE SACT tumour level linkage limitations

- for a SACT record to be automatically linked to a registered tumour, the 3-digit ICD-10 code (primary diagnosis) in SACT must match the registered tumour and the treatment start date must be before the date of death and after the date of diagnosis, within 200 days after the date of diagnosis or within 100 days of another ENCORE diagnosis or treatment event
- if no data has been added to a registered tumour by registration officers for some time eg if the treatment is for recurrence/progression, and no pathology has been received, it is unlikely the SACT record will be linked
- the greater the time between diagnosis and the SACT treatment the lower the chance of the data being automatically linked
- if the rules above are not met, SACT records may be manually linked to a registered tumour by a registration officer. For the manual processing of SACT the following caveat applies:

- SACT treatments submitted after a registration record has been finalised are unlikely to be linked. Additionally, when manual matching, registration officers concentrate on the first six months' worth of SACT data due to the volume of data received

Applicants should consider these limitations when determining whether to request tumour or patient level linkage for any ODR request. The added complexity of the tumour level linkage will likely affect the cost estimate of the data delivery.

Where patient level linkage is requested but the pseudonymised cancer registration tumour ID is also required, all SACT treatments for each patient will be provided, and therefore it should be noted that each SACT treatment will not necessarily relate to the tumour ID provided. However, providing all SACT treatments using patient level linkage enables the applicant to determine themselves the most likely tumour each SACT treatment record is associated with.

Cancer Waiting Times (CWT) treatments

Data is available for first and subsequent cancer treatments recorded in the National Cancer Waiting Times database, regardless of the diagnostic or referral route. All CWT treatment records are algorithmically linked to the most relevant registerable tumour from the cancer registry data.

Approximately 80% of tumours (ICD-10 C00-C97, D05 exc. C44) are linked to at least one treatment record in the National Cancer Waiting Times database. The dataset includes various dates to monitor progression through pathways as well as basic details on treatment and referral information.

Please contact the ODR team and request a conversation with the NCRAS analytical team to discuss any queries you have, after reading the above guidance and before submitting your ODR application.