Public Health England

Protecting and improving the nation's health

A guide to NCRAS data and its availability

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 October 2020
PHE publications
gateway number: GW-1610

PHE supports the UN
Sustainable Development Goals

Contents

Aim

Public Health England's National Disease Registration Service (NDRS) includes the National Cancer Registration and Analysis Service (NCRAS), one of the largest, most advanced and complex cancer data curation service anywhere in the world and the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS). NCRAS collects data on all cases of cancer that occur in people living in England. NCRAS routinely collects both patient and tumour level information to build a rich picture of the cancer patient pathway. The data is used to support public health, healthcare and research. NCRAS holds a large variety of datasets on cancer, some of which are sourced from external organisations (such as Hospital Episode Statistics, HES from NHS Digital), and others which flow directly to PHE from NHS Trusts (such as Cancer Outcomes and Services Dataset, COSD). The PHE Office for Data Release (ODR) is responsible for managing the release of personally identifiable or de-personalised data from PHE. This guide aims to explain the flow of data into NCRAS and the differing availability of data items based on their source.

Cancer registration

Many NHS Trusts use a cancer management system (such as Somerset, Infoflex and others) within their hospital. An export is taken from the cancer management system for the Cancer Outcomes and Services dataset (COSD) which is typically once a month and emailed to the team at PHE. These extracts are then loaded into our internal system, and the process of cancer registration commences.

NCRAS also receive different feeds from different data sources within the Trusts such as Patient Information Systems (PAS), Pathology, Radiotherapy, and E-prescribing. Each of these feeds are often on different systems – PHE maps and links the data for each patient and tumour onto our own internal system.

Cancer registration is a process undertaken by registration officers across the country which ensures the data quality of the information in our system enabling robust analysis of cancer care and outcomes. This process populates a series of tables in the Cancer Analysis System (CAS) known as the "National Cancer Registration Dataset" tables. Undertaking this process requires time and resource, including time to elapse for treatments to occur post-diagnosis, and therefore the cancer record arising from this process is finalised some **12 to 15 months** after initial diagnosis. Further details of this process are available here: https://academic.oup.com/ije/advance-article/doi/10.1093/ije/dyz076/5476570

Availability of linked data

The National Cancer Registration Dataset tables are enhanced with data from other sources by linking at either a patient or tumour level, including data from the Radiotherapy Dataset (RTDS), Systemic Anti-Cancer Therapy Dataset (SACT), HES (inpatient, outpatient and A&E), as well as others (see table below for full list of linked datasets, and their availability). Updates to the linked tables occurs as more data becomes available, and occurs on a table-by-table basis.

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 information on each of these datasets, including the date ranges for which they are available, can be found in the NCRAS data dictionary for ODR requests. The length of time needed for cancer registration has a knock-on impact on when other, linked, datasets are available through NCRAS, as NCRAS' remit to collect data covers information about cancer patients and patients who are suspected to have a cancer diagnosis. This means that for HES (inpatient, outpatient and A&E) and Diagnostic Imaging Dataset (DID) data, we can only analyse linked data once registration is finalised, as the cohort is defined from the National Cancer Registration Dataset tables. The PHE DataLake[1] contains unfiltered HES, however we only use this for metrics which do not use any information from the National Cancer Registration Dataset tables other than the fact of previous diagnoses, such as the Emergency Presentations metrics.

Further details regarding the purpose, structure, and research uses of the Hospital Episode Statistics Admitted Patient Care (HES APC) data, including information on linkage to other datasets such as Cancer Registration are available here: https://www.ncbi.nlm.nih.gov/pubmed/28338941

Further details regarding the purpose, structure, and research uses of the SACT dataset are available here: https://academic.oup.com/ije/advance-article/doi/10.1093/ije/dyz137/5538002

[1] The 'data lake' is a combination of different data sources but the assets are not consistently catalogued or labelled making combining data from multiple sources to answer questions difficult.

The ability to define a specific cancer cohort from stand-alone external datasets is generally much more limited than would be possible if defining a cohort from the National Cancer Registration Dataset tables.

For example, we receive data on the systemic anti-cancer therapy (SACT) treatments patients are receiving. If relying solely on data within the SACT feed and not on that in the National Cancer Registration Dataset tables, the ICD-10 codes of the tumours are known to have several data quality issues, for example clinicians may record a cancer under a different ICD-10 code to be able to select the type of SACT they provided.

Similarly, for the Cancer Waiting Times dataset (CWT), as this data is recorded before a diagnosis of cancer has been made, only a top-level breakdown of types of suspected cancer are possible until the data is linked to the Cancer Registration information sourced from the full information available about the patient and tumour.

The date range of the data available is therefore dependent on the specific analytical question, however the date ranges given in the most recent NCRAS data dictionary for ODR requests outline the data that is available for full analysis.

Other COSD data items held by NCRAS in raw, unstructured form

Whilst the NCRAS data dictionary for ODR requests (available here) is the best place to start to understand which data items NCRAS routinely makes available for further analysis and research (and are typically the variables NCRAS analysts themselves would utilise for analysis). The data items included in the NCRAS data dictionary for ODR requests are not a comprehensive list of all those that flow into PHE.
NCRAS receives additional data items to those in the National Cancer Registration Dataset as part of the COSD data returns, however, many of these are stored in raw, unstructured form, and cannot be easily extracted. NCRAS has developed, tested and finalised algorithms to take raw, unstructured data, which contains duplicated values and/or inconsistent values from multiple trust returns, and reliably extract and convert this into a single value per time point of interest. Once this process, which is complex and time-consuming is complete, items are then made available in the National Cancer Registration Dataset, and consequently the NCRAS data dictionary for ODR requests.

NCRAS is currently undertaking work on many COSD variables to create the algorithms to move them into a more structured form, and specific data items are being

prioritised based on NCRAS' understanding of their utility for further analysis. As this work progresses, items will become available in the National Cancer Registration Dataset, and therefore available for ODR requests.

In addition to the duplication and inconsistency of the raw, unstructured data items, many of these COSD data items have very poor data completeness, so their analytical use may be limited, or there are known issues with data quality, making analysis using such items unreliable. Work is ongoing to improve the reporting of data completeness for clinical users to provide leverage for Trusts to improve their submissions of key data items.

For COSD data items where bespoke algorithms have not yet been developed to enable the robust and reliable extraction of a single value per time point of interest, NCRAS is developing a generic simpler algorithm to extract all data submissions for a given data item, including all duplicates and inconsistent returns. Where an applicant has a specific rationale for requesting this raw, unstructured data, understands the messy and complicated natures of these fields, and has the capacity and analytical capability to develop their own algorithms to be able to interpret the data, these data items can be specifically requested. However, NCRAS will not be able to advise on how to develop the necessary algorithms, or provide any analytical support to achieve this. The applicant should give careful consideration as to whether other additional variables may also need to be requested in order to determine which value of the raw data item is correct. As a general rule, NCRAS does not recommend this approach except in exceptional circumstances.

Overall, this means that there are many data items listed in the COSD dataset which are not sufficiently complete or of robust enough data quality to be used for analysis just yet. **Therefore, items should be selected from the NCRAS data dictionary for ODR requests and not from the COSD dataset.**

Dataset curated by NCRAS

Dataset	Description	In the ODR offer?	Patient or tumour linked?
National Cancer Registration Dataset	NCRAS merges data from multiple datasets to create the National Cancer Registration Dataset. This dataset contains 3 types of tables for patient information, tumour information and treatment information, which are linked by NHS number. The national cancer registration dataset includes a subset of COSD, as well as the Route to Diagnosis for each tumour, Charlson co-morbidity, and information from the Index of Multiple Deprivation (IMD).	Yes	Tumour
Cancer Outcomes and Service Dataset (COSD)	The Cancer Outcomes and Services Dataset (COSD) is the national data standard for reporting cancer in the NHS in England and is collected and managed by NCRAS; it has been mandated/collected since 2013. It is the overarching framework that describes the cancer datasets collected in England. The COSD data structure is extensive, containing 489 items in version 8[2] and covers clinical and pathological items. The data structure specifies the information to be sourced from a number of datasets, such as operational data on patient waiting times, treatment data including surgery, chemotherapy and radiotherapy, and mortality data.	Additional data items to those in the National Cancer Registration Dataset not available as standard – see previous section	Patient
Systemic Anti-Cancer Therapy (SACT) dataset	Since April 2012 NCRAS has collected data on systemic anti-cancer therapy (SACT) activities, which includes chemotherapy, from all NHS England chemotherapy providers to create the SACT dataset. The SACT dataset collects clinical information about treatments given to patients.	Yes	Patient – work ongoing to create tumour-linkage

[2] New versions of COSD are periodically released

Radiotherapy dataset (RTDS)	Introduced in 2009, the Radiotherapy Dataset (RTDS) is the national data standard for collecting information about radiotherapy treatment. The RTDS collects data from all NHS Acute Trust providers of radiotherapy services in England, who submit data to NCRAS monthly.	Yes	Patient
Somatic acquired molecular dataset	NCRAS has been collecting somatic molecular data from 2016 onwards. Somatic tests are performed directly on tumour tissue to identify molecular abnormalities that are specific to the tumour and not present elsewhere in the body. We record all aberration types, from very small variants at the DNA level up to large chromosomal abnormalities. The results of somatic testing are used for cancer diagnosis, prognosis and increasingly for precision medicine to identify the most appropriate treatment for a tumour based on its molecular profile (targeted therapies).	No – work ongoing to quality assure data and will then be included	Tumour
Germline dataset	NCRAS collects germline molecular data performed in individuals with a strong familial predisposition to cancer. Germline testing differs from somatic testing in that the molecular abnormalities detected are not restricted to the tumour but instead are present in every cell in the body, causing a high lifetime risk of developing cancer. Pilot data collection work has focused on the BRCA1 and BRCA2 genes, with NCRAS leading and coordinating the contribution of England and Wales to the BRCA Challenge, an international collaboration to aggregate anonymised data on variants within the BRCA1 and BRCA2 genes on a global scale. The germline work is also being extended to include hereditary colorectal cancer predisposition syndromes and others.	No – long term plan to add	Patient
National Cancer Diagnosis Audit (NCDA)	NCRAS collects and manages data for the National Cancer Diagnosis Audit, which contains information about the diagnosis of cancer patients in primary and secondary care. Participating GPs securely submit information about the primary care part of the pathway for their patients who were diagnosed with cancer during the timeframe selected for the audit.	Yes	Tumour

National Clinical Audits for Lung, Breast and Prostate Cancer	NCRAS manages the data collection for the National Lung Cancer Audit[3], the National Prostate Cancer Audit and the National Audit for Breast Cancer in Older Patients.	Yes	Tumour

[3] More information on the availability of lung cancer data is given in Appendix 1

Datasets accessible through NCRAS

NCRAS can link registration data to additional datasets collected by other organisations such as NHS Digital. This data linkage enables more comprehensive analysis of cancer data for public benefit.

Dataset	Description	In the ODR offer?	Patient or tumour linked?
Hospital Episodes Statistics (HES)	The Hospital Episode Statistics (HES) dataset is collected and managed by NHS Digital and contains administrative data on hospital admissions within the NHS. This includes inpatient, outpatient and A&E attendances and appointments. It is used by the NHS to allow hospitals to be paid for the care they deliver.	Yes	Patient
Diagnostic Imaging Dataset (DID)	The Diagnostic Imaging Dataset (DID) collected by NHS Digital contains information on diagnostic imaging tests carried out by the NHS. It includes details about the type of test used, where on the body it was conducted, the source of referral as well as patient demographic information.	Yes	Patient
Prescription data	The prescriptions data collected and managed by the NHS Business Service Authority (NHSBSA) contains details of the treatments dispensed in primary care from GPs, pharmacists, dentists and opticians. The prescriptions records for cancer patients only can be linked to cancer registration data by NHS numbers to understand more about treatments and investigate patterns in prescriptions before, during and after a diagnosis.	No, planned	Patient
National Cancer Waiting Times (CWT) Monitoring data	The National Cancer Waiting Times (CWT) Monitoring dataset from NHS England includes data on the time between referral, diagnosis and treatment for cancer patients in the NHS. It is used to support the cancer waiting times standards between different points of the cancer pathway.	Yes	Tumour
National Cancer Patient Experience Survey (CPES)	Commissioned by NHS England, the National Cancer Patient Experience Survey (CPES) dataset is based on surveys sent to patients asking about their cancer journey from symptoms, diagnosis and treatment to aftercare.	Yes	Patient

National Head and Neck Cancer Audit	Managed by NHS Digital, the National Head and Neck Cancer Audit collected data from hospitals in England and Wales on the diagnosis and treatment of patients with cancer of the head and neck, covering patients diagnosed up to 2014.	No, planned	Patient
Patient Reported Outcome Measures (PROMs) datasets	Data from Patient Reported Outcome Measures (PROMs) data collections are held by NCRAS. PROMs surveys collect data from patients about their experience of their journey from diagnosis to aftercare. The data provides a snapshot of patient reported outcomes at specific points in time when the surveys were conducted and are not measured and collected routinely by NCRAS. Datasets held by NCRAS include bladder, breast, colorectal, gynaecological, Non-Hodgkin Lymphoma and prostate cancer snapshots.	Yes	Patient

Appendix 1: Lung cancer data availability

Summary

The most comprehensive data for lung cancer cases is contained in the National Cancer Registration Dataset and NCRAS therefore recommends lung cancer cohorts are defined using this dataset. This cohort of cancer patients can then be linked to other datasets including both the historic LUCADA data and the ongoing additional data collection for NLCA to include supplemental information not available in the National Cancer Registration Dataset

Some data items are more complete in the lung cancer audits datasets than in the National Cancer Registration Dataset, and data is also available from the audits for years predating COSD submission (2012). These data items can be obtained by requesting the lung cancer audit data (LUCADA for 2005 to 2014 and NLCA for 2015 onwards)

In certain cases, additional data items which are only available from LUCADA may be required for particular analytical questions, however applicants should note these data items are only available for diagnoses from 2005 to 2014.

Background

The National Lung Cancer Audit was initiated in 2005 and has been hugely influential to improve care of lung cancer patients in England. The data for the audit were initially collected through the National Lung Cancer Audit database (LUCADA) portal. The resulting LUCADA reports were very visible publications in the lung cancer community and it may not be widely known that data for the lung cancer audits is now collected and collated by NCRAS through COSD submissions and used to populate NLCA reports.

The information provided here seeks to outline the changes in data collection for lung cancer over time, and the key information applicants should be aware of when requesting lung cancer data through ODR in order to ensure receipt of the most appropriate data for the intended analyses.

Lung cancer data sources

NCRAS holds data on lung cancer from 2 different sources:

- the LUCADA portal (2005 to 2014)
- cancer registration including COSD data returns from 2012 onwards (including data in the National Cancer Registration Dataset), which was used for the NLCA data collection

LUCADA (2005 to 2014):

From 2005 to 2013 the national lung cancer audit collected data from trusts and populated an external database called LUCADA. LUCADA collected data on both lung and mesothelioma diagnoses (ICD10 C34 and C45). This data is now stored by NCRAS as a separate stand-alone table, and data items from this table are available for request through ODR. As this database was not part of the Cancer Registry, it contains some items which are not collected in COSD data returns. LUCADA did not capture all lung cancer or mesothelioma registrations and therefore if an applicant were to define their cohort of lung or mesothelioma patients from LUCADA data alone, this will miss a significant number of lung and mesothelioma cancer registrations (ICD10 C34 and C45) for those years[4].

The early years of LUCADA in particular had estimated case ascertainment below 60%[5]. In addition, LUCADA defined cancer registrations by date first seen by a trust whilst the cancer registry and the NLCA define by date of diagnosis, therefore the time periods covered by each year of LUCADA are not the same as the later audits (NLCA). It is also important to note that LUCADA was a patient, not tumour-based data collection, and if a patient in LUCADA developed a second primary lung cancer, their record would be overwritten to reflect the most recent information. Therefore, it is possible that the tumour information for a patient in, for example LUCADA2013 is different from, for example LUCADA2005. The version of the LUCADA data available from the ODR for this time period is the LUCADA2013 dataset covering cases with a date first seen from 2005 to 2013.

Data for cases with a date of diagnosis in 2014 are stored in a separate table called LUCADA2014. The data is separate from other years because in 2014 around half of trusts submitted records via the LUCADA audit portal but also submitted information via COSD to the Cancer Registry. The remaining trusts only submitted information via COSD. NCRAS released data from COSD to the audit team to support their audit

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5521232/
[5] https://digital.nhs.uk/data-and-information/publications/statistical/national-lung-cancer-audit

reports. NCRAS has collated both the data sources into a stand-alone table separate from previous years. The LUCADA fields are therefore available for 2014 but are considered incomplete. Data items from this table (LUCADA2014) are also available for request through ODR.

Some data items, such as performance status and clinical nurse specialist, are available from both LUCADA and COSD data returns; requesting these items from both sources will ensure the data span the widest time frame and has the highest data completion for all lung and mesothelioma cancer cases. However, information from different sources may be contradictory, and the applicant will need to resolve any discrepancies in their analysis, the NCRAS analytical team can help advise on this.

Applicants should be aware that only the LUCADA data items which are now also collected through COSD data returns are available for diagnoses after 2014. Therefore, there are some LUCADA data items which are restricted to diagnoses up to 2014, and are only available for the sub-set of lung and mesothelioma cancer patients LUCADA collected data on.

NLCA (2015 to present):

From 2015 onwards, the national lung cancer audit, now called the NLCA, only uses data from COSD data returns. In contrast to LUCADA, the NLCA collects information on lung cancer cases only (ICD10 C34). Each year of data provided to the NLCA audit team to use in their reports (currently 2015, 2016 and 2017) is stored as a stand-alone table by NCRAS, and can be requested through ODR. The NLCA tables include some COSD data items not currently available in the NCRAS data dictionary for all cancers. These items (such as performance status, Clinical Nurse Specialist and MDT discussion date) utilise all COSD sources including raw data and can be requested on the “NLCA” tab within the NCRAS data dictionary for lung cancers only. Only one value per tumour is available for all these data items; this is the value which was used for the audit analysis and further information can be found in the audit reports.

National Cancer Registration Dataset (1995-present):

In contrast, lung and mesothelioma cancer data in the National Cancer Registration Dataset is collated by NCRAS with further information being added to the dataset as it becomes available. Updated snapshots of this dataset are stored and can be requested through ODR.

Table 1. Data item availability

	1995 to 2004	2005 to 2013	2014	2015 to present
LUCADA only data items		X	X (not complete for all trusts)	
COSD NLCA – lung audit specific items collected through COSD such as performance status, Clinical Nurse Specialist and MDT discussion date			X (items collected change over time)	X
National Cancer Registration Dataset data items (including items collected by both LUCADA and COSD)	X	X (some data items can be supplemented with info from LUCADA)	X (some data items can be supplemented with info from LUCADA)	X