Public Health England

Protecting and improving the nation's health

National Cancer Registration and Analysis Service

Guide to using the Simulacrum and submitting code

June 2020

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

For queries relating to this document, please contact: NCRASenquiries@phe.gov.uk

Published June 2020
PHE publications
gateway number: GW-1332

PHE supports the UN
Sustainable Development Goals

Contents

1. Background

The Simulacrum is synthetic cancer data which imitates some of the data held securely by the National Cancer Registration and Analysis Service (NCRAS) within Public Health England (PHE). The Simulacrum looks and feels like the real cancer data held within NCRAS, but does not contain any real patient information. Anyone can use it to learn more about cancer in England without compromising patient privacy. Also, because the Simulacrum data model is the same as the real one in PHE, the Simulacrum can be used to write and test queries that (with the right permissions and ethical approval) could be run on the real data.

Although the data is synthetic, the Simulacrum maintains most of the properties of the original data with a high degree of accuracy. But, there are limitations; the more complex the data query the more approximate the results. Because the data model (but not the data) is the same as the real model in the Cancer Analysis System in PHE, researchers can use the Simulacrum to plan and test their hypotheses before making a formal request to PHE to analyse the real data.

The Simulacrum is entirely synthetic data and is available for anyone to use. Because it only approximates the original data, results from the Simulacrum should not be used for clinical decisions.

The Simulacrum was developed and built by Health Data Insight (HDI) CIC.

The purpose of this document is to provide external researchers, including charities, academics, NHS organisations and industry partners, with a clear understanding of the Simulacrum data and how it can be used to gain an understanding of the structure and content of the cancer data held by NCRAS before making data access requests.

2. Characteristics of the Simulacrum and how it can be used

The Simulacrum contains data about synthetic patients diagnosed between 2013 and 2015, such as age and sex, and data about synthetic tumours, such as staging and pathology information (simulated from the National Cancer Registration Dataset). Like in real life, the synthetic patients can have multiple tumours. The vital status of each synthetic patient up to the end of 2017 has also been simulated so researchers can analyse survival using the Simulacrum data.

Data about treatments with Systemic Anti-Cancer Therapy (SACT), most commonly chemotherapy, has also been simulated. This includes information about the types and number of treatments received. SACT data has been simulated following diagnosis and this has been based on data available up to March 2018. With this data researchers can analyse the treatments following diagnosis.

Simulacrum data cannot be used to answer epidemiological questions as the data is artificial. It is designed to support the preparation of hypotheses and the structuring of questions, so the questions can later be asked using real patient data held by PHE in the Cancer Analysis System (CAS). The Simulacrum structure and data are designed to closely match CAS. This means the Simulacrum data can be used to explore the data, to answer some data quality questions, including determining the data completeness of specific fields, to assess the feasibility of answering specific analytical questions, or as a test dataset for machine learning.

Examples of feasibility questions that the Simulacrum can help answer include:

- “Is a particular SACT drug recorded in CAS?”
- “What are the death codes and morphology codes recorded in CAS for specific cancers?”
- “Can particular cancer cohorts which are not routinely reported on be identified and defined using CAS?”

The Simulacrum can be used to explore data quality questions in more detail than can be answered using the CAS Explorer. It can be used to determine whether particular variables in the dataset have appropriate completeness and distribution to enable specific comparisons for cancer types, or treatment types, or patient cohorts.

Where a researcher is interested in submitting an Office for Data Release (ODR) request, the Simulacrum data can be used to help determine the data items to request, to understand their data completeness, and to help determine the definition of the cohort of tumours the researcher wishes to request data about. This means that the

researcher can come to understand whether the data exists and is complete enough to support analysis, before making a formal request to ODR for the data itself. Where the analytical team are supplied with working and appropriately structured PL/SQL code for an ODR request, this is likely to reduce the time taken to extract the data and fulfil the request, and therefore reduce any associated costs.

The expectation is that most Simulacrum work will come to be run on CAS itself through NCRAS Enquiries or ODR routes.

3. Developing and running code on real PHE data

Once an external researcher has refined their query using the Simulacrum, it is then possible to make a request to Public Health England to have the queries run on the real NCRAS data. The process is outlined in the decision tree in section 6.

External researchers can analyse the Simulacrum using their preferred analytical package. However, as the CAS currently runs on Oracle SQL if the researcher wants to request that Public Health England run their queries on the real NCRAS data, the queries must be created using PL/SQL for data extraction, and R or Stata for analytical modelling using the extracted data. Here are some PL/SQL query examples to illustrate this.

Researchers can develop code on the Simulacrum data either for an aggregate data release or for a depersonalised row-level release. The NCRAS Enquiries inbox can release only anonymous data (i.e. data that meets the ISB Anonymisation Standard). If the external researcher requires data which has a higher risk of being identifiable and which does not meet the ISB Anonymisation Standard (whether aggregate or depersonalised row-level), they will need to make a formal request to the ODR.

It is possible that the exact nature of the data is not known until the query has been run, however, NCRAS Enquiries will only undertake queries on real data where the external researcher has sufficiently demonstrated that the query run on Simulacrum data meets the ISB Anonymisation Standard. The external researcher should therefore assess their query on Simulacrum data to ascertain whether the simulated data meets the ISB Anonymisation Standard. The ISB Anonymisation Standard outlines the standard anonymisation processes for health and social care data to assess the risk of extra information being used to try to reveal the identity of individuals. It includes a set of standard anonymisation plans that can be used to reduce this risk, and to ensure the

release of non-identifying data. For individual-level data, a common anonymisation plan would follow ‘Plan 3’ in the ISB Anonymisation Standard whereby the data are derived to ‘weak’ k-anonymity by reducing the detail in indirect identifiers. The external researcher should supply this draft assessment to NCRAS along with relevant evidence and justification for the anonymisation plan selected. It will then be reviewed by an NCRAS analyst and the NCRAS Caldicott Guardian. If the external researcher is not able to undertake such an assessment themselves to demonstrate that the requested data is anonymous, or the conclusion is challenged by the NCRAS Caldicott, the researcher should instead make either a formal request to the ODR for depersonalised data or should contact simulacrumdata@healthdatainsight.org.uk for more information on the request service HDI is facilitating in partnership with IQVIA.

NCRAS Enquiries are only able to run code for anonymous data releases developed on the Simulacrum if the whole process is expected to take less than 3 hours of analytical work for the NCRAS analytical team to run the code and check the resulting aggregate or depersonalised figures for compliance with the ISB Anonymisation Standard. If unexpected issues arise during the process which means the time taken to complete the request will go over 3 hours, the request will need to be redirected to either HDI or ODR.

NCRAS Enquiries are unfortunately unable to provide additional support to understand the Simulacrum data or any detailed technical advice, however, researchers can contact ODR to discuss the formal process for row-level releases, or can contact HDI (simulacrumdata@healthdatainsight.org.uk) for further help with the Simulacrum. Please note, there will be costs associated with both these options.

NCRAS Enquiries provides a request service for simple analysis, and requests are placed in a queue until an analyst is available to undertake the work. For complex or repeated requests for bespoke analysis, HDI is working in partnership with IQVIA and facilitates a request service through this partnership – contact simulacrumdata@healthdatainsight.org.uk for more information.

4. Technical guidance on using the Simulacrum

Aspects to consider when using the Simulacrum data

Consider the level of data being requested. For example, if a researcher is interested in patients receiving carboplatin, consider that one patient could be diagnosed with multiple cancers tumours in their lifetime and receive multiple rounds of carboplatin treatment for each, therefore the researcher must define whether they wish to count patients, tumours or events.

Consider how to define a particular agent of interest. For example, if a researcher is interested in patients in receipt of Cisplatin, the researcher must determine whether they wish to include all regimens that include Cisplatin, or whether to include only if Cisplatin is a single agent.

Consider patients who have multiple diagnoses. For example, if the researcher is interested in ovarian cancer, consider whether to include or exclude patients who received a second primary cancer diagnosis **before** the ovarian cancer diagnosis, and similarly, whether to include or exclude patients who received a second primary cancer diagnosis **after** the ovarian cancer diagnosis.

Consider the variety of treatments recorded in SACT. For example, there are some non-chemo and hormones regimens recorded in SACT, and these are typically excluded by the NCRAS analytical team during analysis unless they are of specific relevance to the analysis. The researcher must therefore consider whether to include or exclude such treatments.

Consider whether tumour or patient linkage is more appropriate for the research question of interest. Data on each event, treatment or outcome in the simulated SACT tables has been linked to information in the simulated National Cancer Registration Dataset tables at a patient-level using the LINKNUMBER variable. However, this does not differentiate cases where the patient has had more than one cancer after their initial cancer diagnosis. NCRAS has developed an algorithm to tumour-link the real SACT data to the real National Cancer Registration Dataset data which identifies SACT treatments occurring from 31 days before a specific tumour diagnosis, and then further builds on this linkage by typically utilising information on the date of diagnosis, referral or treatment, as well as the cancer site of the tumour, with the aim of identifying the tumour most likely to relate to each event, treatment or outcome (where appropriate). Patient linkage may be more appropriate when analysing cohorts of patients with only a single primary diagnosis on record, whereas tumour linkage may be more appropriate where patients are known to have received more than one primary cancer diagnosis. Obtaining tumour-linked SACT data using the NCRAS algorithm will require an ODR request.

Consider the format of the desired output. For example, is the final output the researcher is aiming for a single number, or a spreadsheet tab with specific columns for Cancer Type, Stage, or separate tabs for patient-level breakdowns and tumour-level breakdowns. Producing template tables to be populated can help guide the development of code needed to create the desired output.

5. SACT information and FAQs

Overview and limitations of SACT data:

The SACT data contains information on patients treated with chemotherapy from April 2012 onwards. However, the SACT data collection was not mandated until April 2014 with most trusts conforming by July 2014, therefore researchers should be aware that the SACT data may not be complete prior to July 2014.

There may be a restriction on the release of data on patients receiving treatments funded through the Cancer Drugs Fund. Each request for data will be checked and restrictions applied if required.

There are known gaps in the reporting of SACT data on the following:

- haematological cancers – likely due to differences in coding and the complexities of treatments
- childhood and young adult cancers – likely due to paediatric treatment regimen complexity
- oral and hormone treatments – likely due to these treatments being prescribed outside of the hospital pharmacy setting, such as in the community or primary care, or recorded on something other than a hospital e-prescribing system
- at a treatment level, completeness is not known – however, it has been found that a high number of regimens contain only a single cycle (around 20%), indicating that not all cycles are being reported for each regimen, or that many cycles are being incorrectly split into separate regimens. All regimens may also not be reported for each patient

How reliable are the data fields ‘START_DATE_OF_REGIMEN’ and ‘START_DATE_OF_CYCLE’?

The start date of a regimen should be near the start date of the first cycle within that regimen. However, it is possible for large gaps to exist. For this reason, some project work may choose to use first cycle start date as the true start date of regimen.

Some records have regimens that start before the date of diagnosis, or a regimen that starts after the start date of the earliest cycle within the regimen. These are data quality issues that should be identified and handled when analysing dates extracted from SACT tables.

Records with a regimen start date before 2012 may be especially unreliable and should be excluded from analyses or else the date information subject to close inspection.

Many patients appear only to receive a single cycle. How can this be explained? How does this impact treatment duration analysis?

CYCLE_NUMBER is in the Cycle table and should be sequentially numbered within each regimen and updated in subsequent data submissions. The START_DATE_OF_CYCLE field is also located in the Cycle table, defined as the date of first drug administration in each cycle. This should correspond to cycle number with logical, sequential cycle numbering following chronologically.

If having to choose between the 2 items, START_DATE_OF_CYCLE is more reliable than CYCLE_NUMBER. Cycle numbers should not be used in isolation to identify and order distinct cycles of treatment. Guidance on counting cycles is in the early stage of development.

Records with cycle start dates before April 2012 should be excluded.

On the issue of patients being listed with only a single distinct cycle, this can be typical for certain cancers such as those in children, teenagers and young adults (CTYA) and in haematological cancers, where treatment tends to be more linear. Here, all administrations are listed within a single cycle.

Aside from some CTYA and haematological cancers, it is more conventional for a regimen to contain multiple cycles of treatment, and for each cycle to contain multiple administrations. However, of all non-trial regimens listed in SACT between 2013 and 2017, roughly 20% contain <2 cycles and <2 administrations. A stratification of this proportion by year and provider indicates improvements over time, with reporting issues experienced by certain providers having been resolved since the inception of SACT.

After analysing the median time to death from the start date of these empty regimens, short survival appears to be an unlikely explanation for the absence of repeated cycles and administrations (a median 277 days). Similarly, they were found to be no more prevalent among CTYA patients than non-CTYA patients.

Remaining hypotheses include:

1. transition of patient treatment to a primary care or community setting where data are not captured
2. temporary or permanent discontinuation of treatment due to toxicity or an adverse event
3. patient choice to end treatment early, or
4. patients being unfit for treatment

Cases are likely to comprise a combination of all 4 factors, with transitions of treatment out of secondary care being the most likely. Of regimens that contained <2 cycles and <2 administrations, by far the most common regimens were hormone therapy, making up just over 15% of the total.

How many cycles of treatment has a patient received?

The SACT dataset has 3 fields which can be used to infer the total number of cycles of treatment a patient has been given, these include start_date_of_cycle, merged_cycle_id and cycle_number. After some data exploration it was found that cycle_number was often unreliable, with double counting, not starting at 1 and skipping numbers. Counting the number of merged_cycle_ids often resulted in overcounting from records which were submitted more than once and were deemed duplicates. Counting the start_date_of_cycle was the most reliable way of inferring the total number of cycles per patient/regimen and this is the recommend approach. However, prior to June 2017 it was not mandatory for trusts to submit start_date_of_cycle, therefore by using this approach you could potentially miss cycles, so it is important to look at your data and decide whether this approach is sufficient for your research question.

6. Decision tree for using the Simulacrum and submitting code to NCRAS