NDRS
NATIONAL DISEASE REGISTRATION SERVICE

NHS England

Guide to using the Simulacrum to support NDRS data requests

September 2023

About the NDRS

The National Disease Registration Service (NDRS) is part of NHS England. Its purpose is to collect, collate and analyse data on patients with cancer, congenital anomalies, and rare diseases. It provides robust surveillance to monitor and detect changes in health and disease in the population. NDRS is a vital resource that helps researchers, healthcare professionals and policy makers make decisions about NHS services and the treatments people receive.

The NDRS includes:

- the National Cancer Registration and Analysis Service (NCRAS) and
- the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS)

Healthcare professionals, researchers and policy makers use data to better understand population health and disease. The data is provided by patients and collected by the NHS as part of their care and support. The NDRS uses the data to help:

- understand cancer, rare diseases, and congenital anomalies
- improve diagnosis
- plan NHS services
- improve treatment
- evaluate policy
- improve genetic counselling

NDRS
NATIONAL DISEASE REGISTRATION SERVICE

National Disease Registration Service
The Leeds Government Hub
7&8 Wellington Place
Leeds
LS1 4AP

NHS England

For queries relating to this document, please contact:
NDRSenquiries@nhs.net

Contents

1. Background

The Simulacrum is synthetic cancer data which imitates some of the data held securely by the National Disease Registration Service (NDRS) within the National Health Service (NHS) England. The data collected by NDRS is held in a database called the Cancer Analysis System (CAS). The Simulacrum looks and feels like the real cancer data held within the CAS 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 schema is the same as the real one in the CAS, the Simulacrum can be used to write and test code to run queries that, with the right permissions and ethical approval, can be run on the real data.

The Simulacrum maintains many of the statistical properties of the original data with a high degree of accuracy, for example, the distributions of individuals variables and correlations between variables in the data. This means that one can run queries on the data and get a preliminary idea of what results would look like if run on the real data. However, there are limitations: the more complex the data query, the more approximate the results. For this reason, results from the Simulacrum should not be used for clinical decision-making.

Instead, Simulacrum can be used by researchers to learn about NDRS data and assess whether it is a suitable data source for their research before requesting data access. Researchers can also plan their analysis and write code using Simulacrum to produce analysis outputs from the data, before making a request for a data release. Because the Simulacrum has the same data schema and structure to NDRS data, this code can be run on the real data to produce real data outputs. Such requests can be made directly to the NDRS analytical team or through NHS England's Data Access Request Service (DARS).

The Simulacrum was developed by Health Data Insight (HDI) CIC in partnership with NDRS. There have been multiple releases of Simulacrum based on different subsets of NDRS data, which are freely available for download from the Simulacrum website. HDI also provide a data request service for bespoke analysis.

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 guidance on how to write and submit code written on Simulacrum data alongside a request for data release. It is an updated version of the previous NCRAS Guide to using the Simulacrum and submitting code.

For more detailed guidance on how to formulate queries on the Simulacrum data, please refer to the Simulacrum User Guide.

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

The Simulacrum contains data about synthetic patients and their tumour diagnoses, treatments and molecular diagnostic tests. There are currently two available versions of Simulacrum. The most recent version, Simulacrum v2.1.0 (released April 2023) contains data on synthetic patients diagnosed between 2016 and 2019, such as age and gender, and data about their 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 has also been simulated so researchers can analyse survival using the Simulacrum data.

Synthetic records for the patient's treatments and somatic (tumour genetic) tests have also been simulated. These include details about the Systemic Anti-Cancer Therapy (SACT) treatments (most commonly chemotherapy), radiotherapy treatments and somatic genomic tests received. With this data, researchers can analyse the treatments following diagnosis and the types of genetic mutations and aberrations seen in tumours.

The Simulacrum preserves structural properties of the real data, such as the data schema, table structures and linkages between tables. It also preserves many of the statistical properties of the real data with a high degree of accuracy, e.g., the statistical distribution of values within each data variable and strong correlations in the data. For example, the Simulacrum largely captures the correlation seen between cancer site and gender, such that breast cancer patients are typically female while lung cancer patients are roughly evenly split between male and female, as seen as in the real data. However, as the statistical properties are only approximately preserved, Simulacrum will not always be reflective of the real data, e.g., we see synthetic male Ovarian cancer patients in Simulacrum in larger numbers than the real data.

Therefore, Simulacrum data should not be used to answer epidemiological questions or make clinical decisions as answers will only be approximate and may not be reflective of answers derived from the real data. Instead, it can be used to support the preparation of hypotheses and the structuring of questions, so the questions can later be asked using real patient data on the CAS. We outline three main use cases for researchers:

1. Learn about NDRS data through data exploration.
2. Run preliminary analysis and feasibility tests.
3. Write and test code for analysis.

Learn about NDRS data

Researchers can explore Simulacrum data to learn more about NDRS data, for example, its table structures, table linkages and the information included in NDRS data variables. This provides a more detailed view than looking at the available data dictionaries online.

It's worth noting that, since each version of Simulacrum is based on a snapshot of the CAS, the data structure can differ from the newest snapshot of CAS due to changes over time. This may include things like changes to table structure, data variable names and values. These differences are likely to be small can be easily adjusted for when making data requests. For information where current differences exist, please get in touch with HDI at simulacrum@healthdatainsight.org.uk.

Preliminary analysis and feasibility tests

Researchers can also run feasibility tests on the Simulacrum to see whether NDRS data is a suitable resource for their research. Examples of feasibility questions that the Simulacrum can help to answer include:

- What is the completeness of a data variable of interest?
- Is a particular SACT drug or somatic genomic test recorded in CAS?
- What are the morphology codes recorded in CAS for specific cancers?
- Can cancer cohorts not routinely reported on be defined in CAS?

Such information can be useful for researchers in advance of applying for access to NDRS data through DARS. It can help them to decide if NDRS data is suitable for their research and then to plan their DARS application. By answering such questions, it is possible to decide which data items are needed and define the patient cohort of interest for analysis.

Researchers can also run preliminary analyses to get an approximate idea of what analysis results might look like if run on NDRS data. For "simple" analysis results are quite accurate when compared to the real data. However, the more "complex" the analysis, the more approximate results are. For this reason, analysis on Simulacrum should never be used for clinical decision-making.

Write and test code for analysis

Researchers can also use the Simulacrum to write and test code to run analysis. Due to similar data structure to NDRS data, once researchers have written their code and refined their queries, they can request that the code is run on the CAS data to produce real results. For minor differences between the Simulacrum and newest CAS data snapshot, code can be adjusted by the analyst processing the request. With the right ethical and legal requirements, these results can then be released.

Section 3 outlines the process for submitting code, while Section 4 provides advice on writing code that fulfils submission requirements. For guidance on formulating analysis to produce optimal outputs from CAS data, please refer to the Simulacrum User Guide.

3. Submitting code to NDRS for a data release

Once an external researcher has written and tested code to run analyses using the Simulacrum, it is possible to request that their code is run on the real NDRS data. Requests can be submitted to the NDRS analytical team, DARS or HDI. This section outlines the processes for making such a request. These are also outlined in the decision tree diagram in Section 5.

Generally, requests for simple analyses that produce anonymous data outputs can be made directly to the NDRS analytical team by emailing the NDRS Enquiries inbox, at NHSdigital.ndrsanalysis@nhs.net. If it takes less than 3 hours of a NDRS analyst's time to process the request and produces anonymous outputs, the work can be done free of charge and results are released to the researcher. If the researcher supplies suitable code alongside the request, this can reduce the time needed to run the analysis and increase the likelihood of a data release through this route. Requests are placed in a queue until an analyst from the NDRS analytical team is available to undertake the work.

If the request is expected to take over 3 hours, takes over 3 hours due to unexpected issues or the results are found not to be anonymous through assessments by the NDRS analyst, the request will need to be redirected to NHS England data release services, i.e., DARS. To discuss the formal process for releases of row-level data, researchers can contact DARS directly.

For complex or repeated requests for bespoke analysis, contact HDI to enquire about their request services at simulacrum@healthdatainsight.org.uk.

Any code submitted alongside a request for data release should fulfil certain requirements to ensure that it is as straightforward as possible for the NDRS analyst undertaking a request. Guidance on these requirements and how best to write code are outlined in Section 4.

4. Developing code using Simulacrum for a data release request

This section provides guidance on how best to develop code using Simulacrum to supplement requests to the NDRS analytical team for an anonymous data release. It also describes alternative routes for data release where code does not produce anonymous data outputs. These are also outlined in the decision tree diagram in Section 5.

The NDRS analytical team are unfortunately unable to provide support to understand the Simulacrum data or any detailed technical advice. For help with the formulation of analysis on the Simulacrum, please refer to the Simulacrum User Guide or get in contact with HDI at simulacrum@healthdatainsight.org.uk. The Simulacrum User Guide provides examples of data queries, advice on how to link tables and some considerations that should be made when querying the data, including some data quality aspects. For specific examples of SQL queries on the CAS data please refer to the NCRAS SQL query guide.

When processing a request supplemented by Simulacrum code, a NDRS analyst will need to interpret the code, run the code on the real data and then apply quality assurance of outputs for validity and privacy. Therefore, the researcher should ensure that the code is easily useable by the NDRS analyst, i.e., the code should:

- be written in an appropriate analytical programming language,
- be clearly structured and interpretable,
- produce anonymous data outputs.

Programming language

While researchers can analyse the Simulacrum using their preferred analytical package, if they want to request that NDRS run their code on the real NDRS data, the code must be written in an appropriate programming language that is currently used by the NDRS analytical team.

Since the NDRS data is held in an Oracle SQL database called the Cancer Analysis System (CAS), code for data extraction from the CAS must be written in PL/SQL language. For analytical modelling of data that is already extracted from the CAS, code must be written in R or Stata, which are the standard languages that NDRS analysts use.

Structured and interpretable code

To minimise the time taken for the overall process, the researcher should prepare code that is error-free, clearly set out and interpretable, e.g., by commenting code and providing an analytical plan. This will allow for NDRS analysts to easily understand and adapt the code to extract data from the database and run data analysis.

Anonymous outputs

The researcher submitting code should ensure, to the best of their ability, that the data outputs produced will be anonymous and therefore releasable by the NDRS analytical team.

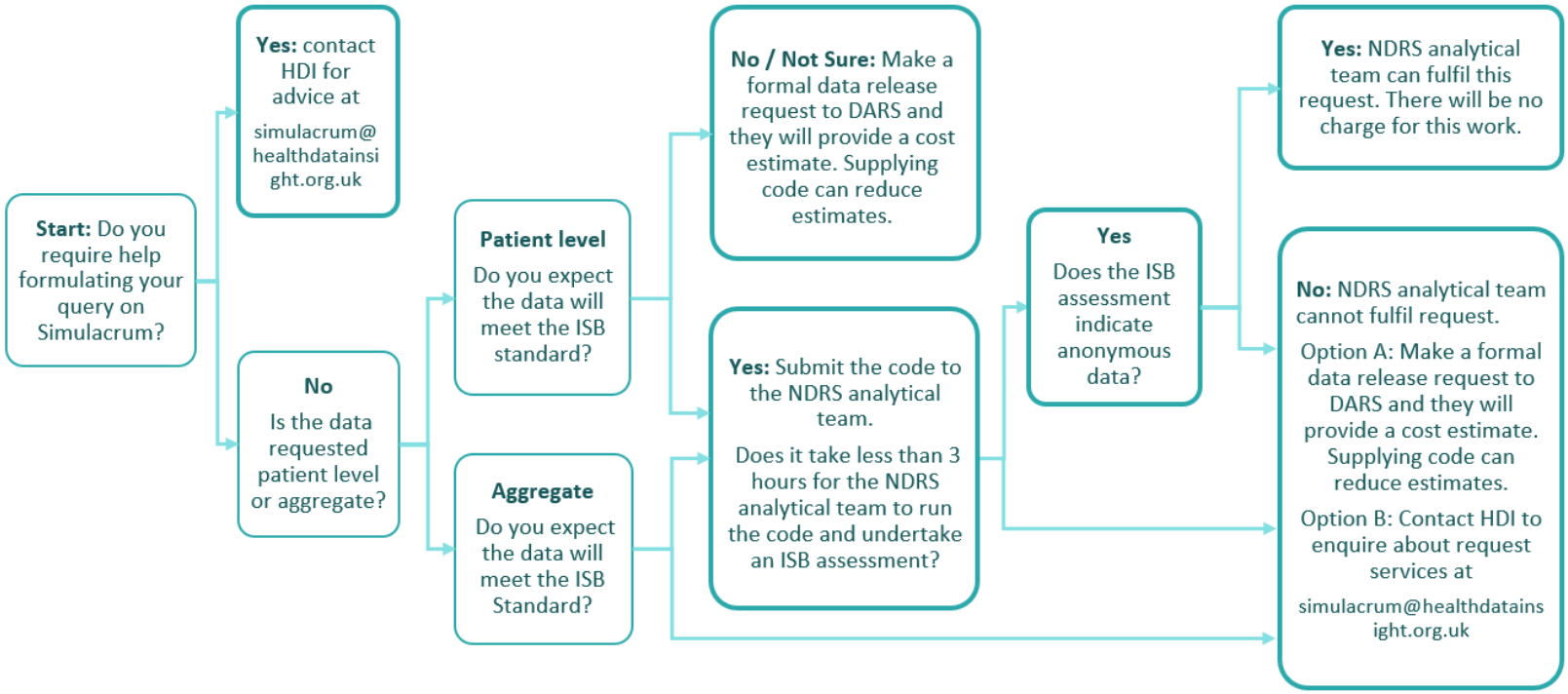
Anonymity assessments should be done according the ISB Anonymisation Standard, which describes the standard anonymisation processes for health and social care data and how to assess the risk of extra information being used to try to reveal the identity of individuals. It includes a set of standard anonymisation plans that can be used to reduce this risk and to ensure the release of non-identifying data. The plan used depends on the type of data outputs being released, which fall into two categories: aggregate data or individual level data. For example, for individual level data, a common anonymisation plan would follow "Plan 3" whereby the data are derived to "weak" k-anonymity by reducing the detail in indirect identifiers. If the code produces NDRS data outputs that do not pass the standard for anonymity, then the request will be rejected or require statistical disclosure control to be applied before release and may be re-directed to DARS.

It is possible that the exact nature of the data outputs is not known until the code has been run on the real data, so making the assessment can be difficult. However, the NDRS analytical team will only accept requests where the external researcher has sufficiently demonstrated that the data outputs are likely to meet the ISB Anonymisation Standard, e.g., by inspecting outputs produced when the code is run on Simulacrum data.

The researcher should supply their draft anonymity assessment to NDRS along with relevant evidence and justification for the anonymisation plan selected. It will then be reviewed by an NDRS analyst and the NDRS Caldicott Guardian. If the 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 NDRS Caldicott Guardian, the researcher should instead make either a formal data release request to DARS or should contact HDI at simulacrum@healthdatainsight.org.uk to explore alternative routes.

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 DARS.

5. Decision tree for using the Simulacrum and submitting code to NDRS



Decision tree: Outlines the process for using Simulacrum to write code and then submit it alongside a request for a data release. The route depends on whether data outputs produced by the code are found anonymous according to the ISB Anonymisation Standard (see Section 4 for more details) and how long the data request takes to process (see Section 3 for more details).