Simulacrum v2 User Guide

Authors: Lora Frayling, Sophie Jose

Date: 21/04/2023

This document supports users of Simulacrum v2 to write code that can be run on the real CAS data.

Contents

Introduction 2

Data Structure 2

Using the Simulacrum 3

Characteristics of the Simulacrum 4

Simulacrum is created from a subset of the available CAS data 4

Simple queries have highly accurate results, while complex queries are less accurate 4

General properties 4

Time series properties 5

Writing code to analyse CAS data 5

NCRD 5

Description of the tables 5

CAS Data quality 6

Guidance on analyses 6

SACT 6

Description of the tables 6

CAS Data Quality 7

Guidance on analyses 8

RTDS 9

Description of tables 9

CAS Data quality 10

Guidance on analyses 10

Genomics Testing Data 11

Description of table 11

CAS Data quality..11

Guidance on analyses: Sample analyses in SQ...11

Introduction

The Simulacrum is a synthetic version of datasets held on the Cancer Analysis System (CAS), collected by the National Disease Registration Service (NDRS) at NHS England. The Simulacrum has similar data structure and statistical properties to the real cancer data, however, does not contain any real patient information. It can be used to conduct data exploration, scope research hypotheses and write and develop code that can be run on CAS data to produce analysis outputs. There have been two versions of Simulacrum, v1 and v2. This guidance document covers all synthetic datasets included in the Simulacrum v2 but is also relevant for those datasets also in Simulacrum v1.

The purpose of this document is to support users of the Simulacrum in the development of such code by providing guidance on:

- The structure of Simulacrum and underlying CAS data
- The properties of the Simulacrum and where they might differ from those of the CAS data
- The writing of code that will produce reliable outputs when run on the CAS data.

For each dataset, it gives:

- Descriptions of each table in the dataset
- The data quality of the dataset in the CAS
- The caveats of the synthetic version of the dataset
- Examples of SQL queries that can be used for analysis.

Data Structure

The Simulacrum v2 is made up of synthetic versions of the following datasets, held on the CAS:

- National Cancer Registration Dataset (NCRD)
    - Information about all patients diagnosed with a cancer in years 2016-2019 and their tumour diagnoses
    - Patient and tumour tables
- Systemic Anti-Cancer Therapy (SACT) datasets
    - Information about all systemic anti-cancer therapy treatments received by patients
    - Regimen, cycle, drug detail and outcome tables
- Radiotherapy Dataset (RTDS)
    - Information about all radiotherapy treatments received by patients
    - Prescription and exposure tables
- Somatic genetic testing data
    - Information about somatic tests performed on patients' tumours

- Gene table.

Simulacrum v1 only comprises data from the NCRD and SACT datasets.

Prior to the generation of the Simulacrum v2, the structure of underlying CAS datasets was simplified to enable easier analysis by users. The resulting data tables all link together to form the relational database. Any code written for analysis on Simulacrum can be run on the equivalent simplified version of underlying CAS datasets. The linkage between tables is depicted in Fig. 1.

There are also other datasets that one can link to on the CAS, e.g. the Hospital Episodes Statistics dataset (HES), however guidance for writing code to do this is out of the scope of this document, as they are not included in the Simulacrum.

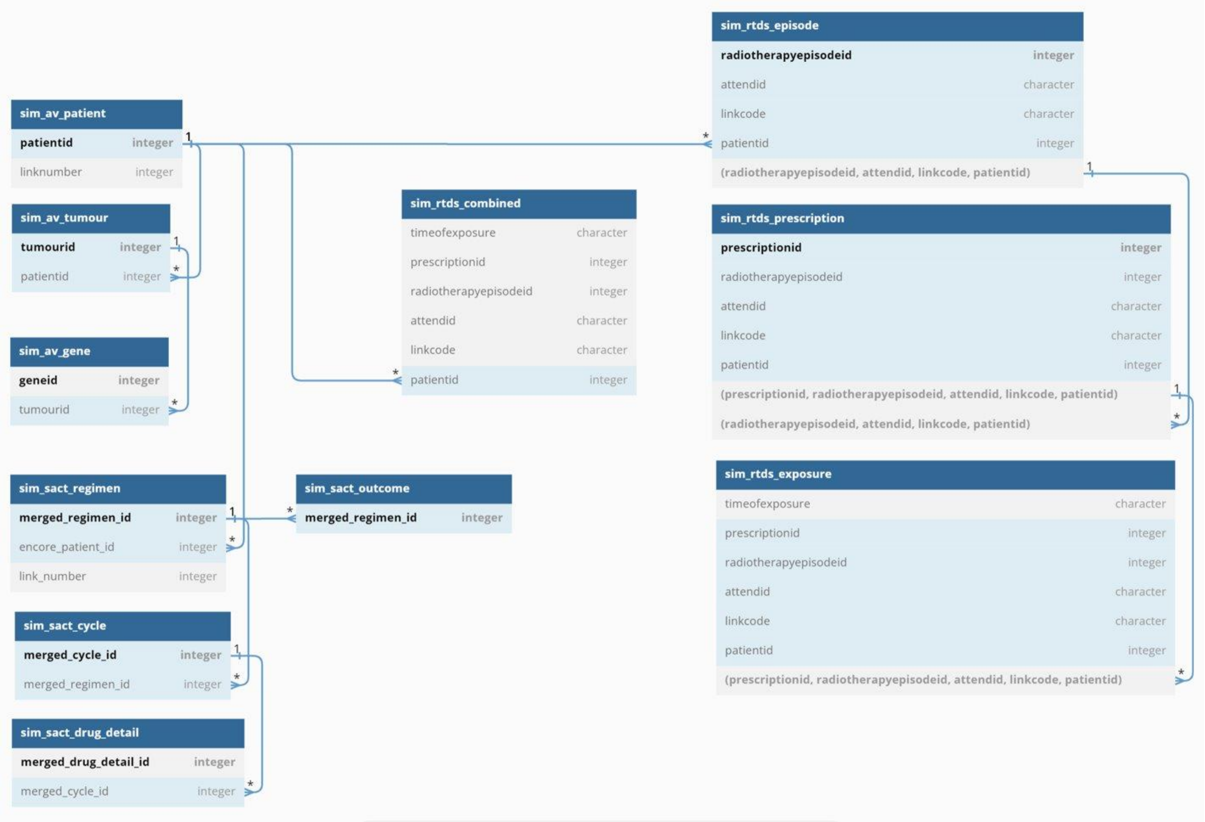

Fig. 1: Table linkage in Simulacrum v2 and CAS (with simplified data structure).

Using the Simulacrum

While the Simulacrum captures many of the statistical properties of the real data, some analyses run on the Simulacrum may look significantly different from the same analyses run on the real CAS data. This is because the synthetic data generation model only captures high-level statistical properties of the real data and the generation model requires that some of these properties are distorted or removed in order to protect patient privacy. Thus, analysis outputs from the Simulacrum cannot be used to answer epidemiological questions or inform real-world clinical decisions.

Instead, the Simulacrum can be used to support the following:

- Initial data exploration:
    - Learning about the data structure
    - Answering data quality questions, e.g. completeness of specific data fields
    - Assessing the feasibility of answering specific analytical questions
- Formulation/refinement of research questions and hypotheses
- Writing and development code to produce analysis outputs, which can then be run on CAS data.

It’s important to understand where differences between analysis outputs run on Simulacrum and CAS may exist as this can affect decisions made on the formulation of research questions, hypotheses and final analysis queries run on the real data. This section provides some guidance on where these differences are likely to occur.

Characteristics of the Simulacrum

Simulacrum is created from a subset of the available CAS data

The Simulacrum v2 currently contains:

- Patients diagnosed with a tumours between 2016 and 2019, whereas the real CAS data contains tumours diagnosed between 1971 and 2020.

Simple queries have highly accurate results, while complex queries are less accurate

- Counting patients with a single characteristic, e.g. the overall number of patients with breast cancer, will produce numbers that are highly representative of the underlying data.
- This becomes less representative as you:
    - Add more characteristics to the patient cohort definition, in particular when linking to SACT, RTDS or gene tables
    - Look at patients with rarer cancer diagnoses.
- Overall, the more complex the analysis being done using the Simulacrum, the less representative we expect the results to be of the same analysis conducted on the CAS data

General properties

- Some of the data variables can have impossible combinations. Examples of this are:
    - In the SACT regimen table, benchmark_group and mapped_regimen do not always match, whereas in the real data they are very closely aligned
    - Sequential heights and weights over time in the SACT data may have impossible combinations

  - Males with gynaecological cancers or females with prostate and testicular cancers occur with higher frequency in the synthetic data than in the real data
- There are fewer one-to-many linkages between tables in the Simulacrum compared to the real data. For example, fewer radiotherapy exposures are simulated per prescription in this dataset.
- There are fewer events than expected in the SACT, RTDS and gene tables. For example:
  - There is an under-representation of genetic tests in 2019 in the synthetic data. This is because the simulation algorithm did not accurately capture an increase in the number of tests arising from an additional data flow from pathology labs for PD-L1 and MMR gene tests.
  - The disparity increases for tables later down the chains of linked tables, e.g. drug detail and exposure tables.

Time series properties

- The Simulacrum v2 method only learns short-term time series patterns, thus long-term patterns will not always be representative of the real data
- Patient records with activity, such as treatments and genetic tests, after a date of death occur at a higher frequency in the synthetic data than in the real data
- The time lengths between events in patients are not always well captured, with a tendency for treatment and gene events to occur closer to diagnosis than in the real data.

Writing code to analyse CAS data

In this section we give guidance on how to formulate analyses that will be run on synthetic datasets in the Simulacrum, and, if needed, the equivalent real datasets in the CAS. For each dataset, we give a description of the tables, an overview of the data quality in the CAS and guidance on how to formulate analyses to produce reliable outputs.

NCRD

Description of the tables

The Simulacrum includes synthetic version of the patient and tumour tables in the NCRD. A profile exists for the NCRD tables and we recommend all users familiarise themselves with this before conducting analysis (https://academic.oup.com/ije/article/49/1/16/5476570). The patient table, *av_patient*, contains patient-level information such as ethnicity, gender and vital status. Each row represents a single patient with the unique identifier PATIENTID.

The tumour table, *av_tumour*, contains tumour-level information related to cancer diagnoses, such as cancer site, stage and date of diagnosis. It includes only primary cancers, not secondary metastases or recurrences. Each row represents a single cancer tumour with the unique identifier TUMOURID. There can be multiple tumour records per patient and each cancer tumour is linked to its corresponding patient via PATIENTID.

CAS Data quality

Information regarding the completeness of data fields in the NCRD can be found on an online tool developed by HDI called the CAS explorer (https://www.cancerdata.nhs.uk/explorer).

Accurately de-duplicated cancer cases are only available from 1995 so it's recommended to only use data from this date for most analyses.

The diagnosis date recorded in the data may not reflect the date that a patient is made aware they have cancer, since histological or cytological confirmation of cancer is prioritised over other sources of information.

Guidance on analyses

For many analyses, the tumour table in the NCRD will be the starting point to define cohorts based on tumour information. Other datasets may then be joined to as appropriate.

Cancers are all coded to ICD-10/O2, both revision 0 and revision 4 (released in 2010), regardless of the coding system in use at time of registration, to enable longitudinal analyses. For revision 0, the SITE_ICD10_CODE and SITE_ICD10_CODE_3CHAR variable can be used. For revision 4, when analysing tumours diagnosed from 2013 onwards (including all tumours in Simulacrum), then SITE_ICD10R4_CODE_FROM2013 and SITE_ICD10R4_O2_3CHAR_FROM2013 should be used. Alternatively, if writing code that will analyse tumours diagnosed before 2013, then SITE_ICD10_O2_PRE2013 and SITE_ICD10_O2_3CHAR_PRE2013 should be used (these are not included in Simulacrum). Cancer site and morphologies are also coded in ICD-O3.

Cancers flagged as being diagnosed on the death certificate (DCO) should be excluded from analyses that assess long term follow-up and outcomes.

Linkage to SACT and RTDS in the simplified CAS data is done via PATIENTID. Since linkage is done at the patient level, one needs consider whether information in these datasets relates to particular cancer diagnoses of interest for study where patients have multiple diagnoses.

Linkage to the gene table is done via TUMOURID.

SACT

Description of the tables

Patients that receive systemic anti-cancer therapy treatments are recorded in the SACT dataset. A detailed description of the SACT dataset can be found in the dataset profile: https://academic.oup.com/ije/article/49/1/15/5538002

Regimens (*regimen*)

Each SACT patient is assigned to a predetermined treatment plan that comprises at least one regimen, recorded in the *regimen* table. A regimen is a single drug or a combination of different drugs given over a planned schedule for systemic anti-cancer therapy treatments and can include chemotherapy and hormone treatments. Regimens link to patients in the *av_patient* table via the PATIENTID.

Cycles (*cycle*)

Typically, a regimen will be given across multiple cycles, recorded in the *cycle* table, reflecting the frequency with which the course of treatment is repeated and the rest periods between administrations. Cycles link to the *regimen* table by MERGED_REGIMEN_ID.

Drug details (*drug_detail)*

Each administered drug within a cycle will have a specific dose and method of delivery (e.g. oral or intravenous), recorded in the *drug detail* table. Drug administrations can be linked to the *cycle* table by MERGED_CYCLE_ID.

Outcomes (*outcome)*

A patient should also have an outcome recorded for each regimen that has been completed or changed (e.g. modified or stopped early), in the *outcome* table. Outcomes link to the *regimen* table by MERGED_REGIMEN_ID.

General

A patient will often have one regimen per line of treatment, with multiple cycles recorded for the regimen and multiple drug administrations per cycle. However, this may not be the case. For example, it is known that (i) some oral drugs are incorrectly prescribed as a single cycle for the duration of treatment, and (ii) cycles for some regimens can become split and submitted as multiple different regimens. This latter issue is understood to be more pronounced in earlier years prior to trusts adopting robust e-prescribing systems.

CAS Data Quality

The earliest data recommended for use is that dated from April 2013 onwards, representing the point from which all trusts regularly submitted the majority of their SACT data.

At a patient level there are known gaps in reporting in the dataset. Three major areas are:

- Haematological cancers
- Childhood/young adult cancers
- Oral chemotherapy and hormonal treatments

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. 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.

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.

Please note that outcomes are not well recorded in the SACT dataset, resulting in low data completeness and quality overall.

Guidance on analyses

Multiple tumours – patient/tumour linkage

Consider whether tumour or patient linkage is more appropriate for the research question of interest. Data on each event, treatment or outcome in the SACT tables are linked to information in the NCRD tables at a patient-level. However, this does not differentiate cases where the patient has had more than one cancer after their initial cancer diagnosis.

Options for tumour linkage to events:

- Removing patients which have more than one diagnosis (the only completely sure way to link tumours to related events)
- Trying to identify tumours most likely to relate to each event, treatment or outcome by utilising information on the date of diagnosis, referral or treatment, as well as the cancer site of the tumour, e.g. link a tumour to a treatment event if the treatment begins within a specific timeframe of date of the tumour diagnosis

NCRAS tumour-link algorithm. This is currently not available through the Simulacrum. Please speak to HDI / NCRAS for more information about this.

Specifying start date of regimen

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 in the data. For this reason, some project work may choose to use first cycle start date as the true start date of regimen.

Counting number of cycles for a regimen

If having to choose between the two items, START_DATE_OF_CYCLE is more reliable than CYCLE_NUMBER for counting the number of cycles in a regimen. Cycle numbers should not be used in isolation to identify and order distinct cycles of treatment due to their unreliability.

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.

SACT versions and data standard

The data standard that Trusts use to submit SACT data has changed over time. This has resulted in changes to some data fields, such as INTENT_OF_TREATMENT and REGIMEN_OUTCOME_SUMMARY. The current dataset standard is SACT v3.0 and all Trusts have been required to submit data according to this standard since December 2019. More information on the current standard can be found at https://www.datadictionary.nhs.uk/data_sets/clinical_data_sets/systemic_anti-cancer_therapy_data_set.html

RTDS

Description of tables

Patients who have received radiotherapy will be recorded in the RTDS dataset.

Episodes (rtds_episode)

An episode is a continuous period of care for radiotherapy including all preparation, planning and delivery of radiotherapy. Each separate episode of care will be entered in the *rtds_episode* table and delivered over one of more appointments. Appointments have their own ID, ATTENDID, and their dates are defined by APPTDATE. In this table there will be one data row per unique episode-appointment combination. Most episodes of care will involve a single prescription of radiotherapy, however, it is possible for there to be more than one prescription per episode. Episodes are linked the patients in the *av_patient* table via the PATIENTID.

Prescriptions (rtds_prescription)

A radiotherapy prescription relates to a series of identical doses of treatments to a single anatomical site. Each prescription is recorded in the *rtds_prescription* table. One prescription will be delivered over the course of one or more appointments, with one data row per unique prescription-appointment combination. Prescriptions can be for either for Teletherapy or Brachytherapy treatment. The *rtds_prescription* table includes information on the anatomical site that received radiotherapy treatment if a patient has metastatic disease, the radiotherapy dose and the treatment region.

Prescriptions can be linked directly to the *av_patient* table by the PATIENTID. The also link to ther corresponding episodes in the *rtds_episode* table via the RADIOTHERAPYEPISODEID, ATTENDID, LINKCODE and PATIENTID.

Exposures (rtds_exposure)

At each appointment for a prescription, patients receive a number of treatment exposures, where each exposure is the delivery of one radiotherapy beam. Exposures are recorded in the exposures table, *rtds_exposure*, with one data row per exposure. This could involve non-treatment exposures which are where the machine is switched on for treatment verification or dosimetry.

Exposures can be linked directly to the *av_patient* table by the PATIENTID. There is a one-to-many link between the prescriptions table and exposures tables, i.e. for each prescriptions table there can be many exposures which can occur on multiple dates at different times. This linkage is via PRESCRIPTIONID, RADIOTHERAPYEPISODEID, ATTENDID, LINKCODE and PATIENTID.

Combined (rtds_combined)

Due to the complex nature of the linkage required between tables within the RTDS dataset, in the Simulacrum we have also provided a combined dataset in which the *rtds_episode*, *rtds_prescription* and *rtds_exposure* tables have already been joined. This table has one data row per exposure and contains all variables from across all the RTDS tables. Users may find it easier to conduct their analysis on this table, particularly if intending to request for queries to be run on the real data in the CAS. If users only require prescription-level information, this dataset can still be used by retaining only distinct records on the subset of variables in the *rtds_episode* and *rtds_prescription* data tables.

CAS Data quality

A few examples of data quality issues are listed below.

- Brachytherapy data is incomplete.
- RTTREATMENTMODALITY is currently not defined correctly as a small number of teletherapy episodes are being classified as brachytherapy. In the meantime, please contact HDI or NCRAS if conducting analysis for brachytherapy or SRS/SRT treatments.
- RTTREATMENTREGION is incomplete for a number of Trusts.
- RTACTUALDOSE and RTACTUALFRACTIONS are incomplete. These fields should be used with caution. It is recommended to use RTPRESCRIBEDDOSE and PRESCRIBEDFRACTIONS over ACTUAL, and to seek clinical advice prior to analysing dose fractionation.

Guidance on analyses

Multiple tumours – patient/tumour linkage

Consider whether tumour or patient linkage is more appropriate for the research question of interest. Data on each event, treatment or outcome in the RTDS tables are linked to information in the NCRD tables at a patient-level. However, this does not differentiate cases where the patient has had more than one cancer after their initial cancer diagnosis.

Options for tumour linkage to events:

- Removing patients which have more than one diagnosis (the only completely sure way to link tumours to events)
- Trying to identify tumours most likely to relate to each event, treatment or outcome by utilising information on the date of diagnosis, referral or treatment, as well as the cancer site of the tumour, e.g. link a tumour to a treatment event if the treatment begins within a specific timeframe of date of the tumour diagnosis

Counting prescriptions

One can count the number of radiotherapy prescriptions for a patient by counting the number of distinct prescriptions records for that patient in *rtds_prescription*.

Counting attendances

One can count the number of appointments a patient attended for a single radiotherapy prescription by counting the number of distinct dates on which the patient received an exposure, i.e. APPTDATE, for that prescription in *rtds_prescription*.

Genomics Testing Data

Description of table

The genomic testing data table, *av_gene*, contains information about molecular tests for somatic (acquired) changes in a tumour's genetic material at gene level. Each row represents a unique combination of TUMOURID and a tested gene/chromosome. Although multiple tests may be performed on a tumour gene, each gene will appear only once per tumour in the *av_gene* table.

The table includes information on the type and date of test and test results for each gene tested. Where multiple tests have occurred, a hierarchy has been applied to report the test with the earliest most definitive result. Tested genes in *av_gene* are linked to tumours in *av_tumour* by TUMOURID.

CAS Data quality

While data is received from the largest laboratory (Birmingham) and 10 other laboratories, data is not received from all the laboratories carrying out genomic diagnostics. This means that data received from genomics laboratories does not have full national coverage. It is therefore important to note that the collected data cannot use all cancer registrations as its denominator.

In addition to somatic tests carried out in designated genomics diagnostics labs, a number of tests are carried out by pathology labs, for which some data is available. Specifically for PD-L1 and MMR somatic tests, data is additionally received from pathology labs for tests conducted in 2019 and is believed to have national coverage. Somatic test data are collected by year that the genetic tests were carried out, and therefore this does not necessarily correspond to the diagnosis date of the tumour for which the test is carried out.

Guidance on analyses: Sample analyses in SQ

To investigate the number of tumours tested for each gene type and tumour site:

```
SELECT gene_desc, and site_icd10r4_o2_3char_from2013, overall_ts, COUNT (DISTINCT
TUMOURID)
FROM gene
GROUP BY gene_desc, site_icd10_o2_3char, overall_ts
ORDER BY COUNT (DISTINCT TUMOURID) DESC;
```

To investigate number of unique genes by overall test result and type of genetic aberration (in this example abnormal genetic aberration):

```
SELECT overall_ts, abnormal_gat, COUNT (*)
FROM gene
GROUP BY overall_ts, abnormal_gat
```

```
ORDER BY overall_ts, abnormal_gat;
```

To investigate number of test dates per gene (this may be useful for tracking resistance to first line targeted therapies through pathway analyses):

```
SELECT count_date, COUNT (*)
FROM gene
GROUP BY count_date
ORDER BY count_date;
```

To link to *av_tumour*:

```
SELECT DISTINCT at.tumourid
FROM gene ge
LEFT OUTER JOIN av_tumour at
ON at.tumourid = ge.tumourid;
```

To count the number of patients who received different regimens and have a specific gene tested, by linking to SACT (e.g. have the gene EGFR tested):

```
SELECT COUNT(DISTINCT atg.patientid), s.benchmark_group, atg.abnormal_gat, atg.overall_ts
FROM gene atg
INNER JOIN sact.at_regimen_england s ON s.patientid = atg.patientid
WHERE s.start_date_of_regimen >= atg.min_date
AND atg.gene_desc = 'EGFR'
GROUP BY s.benchmark_group, atg.abnormal_gat, atg.overall_ts
ORDER BY COUNT(DISTINCT atg.patientid) DESC;
```