Generating the Simulacrum
A methodology overview

Report author: Lora Frayling

November 2018

Introduction

The Simulacrum is synthetic data which has been created in way that mimics some of the data within the Cancer Analysis System (CAS) database in the National Cancer Registration and Analysis Service (NCRAS), which is part of Public Health England (PHE).

This report summarises the methods used to generate the synthetic data. It outlines the project aims, the key methods and steps taken and provides some examples, figures and tables of the methods and results.

A more detailed formal publication will be available in the future; this overview is provided as background to the current release.

For specific technical questions please contact: *simulacrumdata@healthdatainsight.org.uk*

Overview

The synthetic data consists of 7 linked datasets, in the form of data tables that have the same data structure as the data held in the CAS. The current synthetic dataset was modelled on the following two datasets:

- Patient and tumour tables from the Cancer Outcomes and Services Dataset (COSD)
- Patient, tumour, regimen, cycle, drug details, and outcome tables from the Systemic Anti-Cancer Therapy (SACT) dataset.

For example, in the tumour table from COSD, each row corresponds to a tumour diagnosis. Such a row is a detailed record with a set of patient characteristics such as a patient ID (linking the tumour information to the patient's information in the patient table) and the tumour characteristics such as the cancer site, stage, behaviour and other related information. The characteristics in the records form the columns of the table.

Simulacrum data mimics CAS datasets by generating records for a virtual patient population, with no connection to real patients, while preserving the statistical properties of the data in real patient records.

Each Simulacrum dataset is created through a three-step process:

1. Identifying statistical properties of the real data
2. Creating the generative data model
3. Generating the synthetic data

We applied the steps to the anonymised dataset obtained from PHE via the *Office for Data Release* (ODR). Examples of these data cubes can be accessed from: *data.gov.uk*.

Methodology

Our aim for each Simulacrum data table was to capture the cancer-site specific statistical properties. To do this we first split the table records into groups defined by the tumour's cancer site and then applied each of the above steps to each group separately.

Then, within each of these groups we split the table records further into smaller groups in order to capture more site and treatment specific statistical properties. In the Simulacrum COSD tumour table we first split the table records data up into groups defined by the tumour's cancer site. We call these groups *strata*. Each stratum is defined by either a single cancer site or a group of cancer sites, if the numbers are too small to be reported individually. Each stratum has a minimum number of patient records of 1000. Similarly, in the Simulacrum SACT tables, the aim was to capture the treatment-specific properties, so the table records were grouped in strata defined by the patient's treatment.

All of the steps involve methods that require the data to be in a specific form: each characteristic must take one of a set of discrete values, which we call categories. For example, a cancer site takes values from the list of ICD10 cancer codes and each code represents a category.

Methods

Once we have the strata we then applied the following 3-step methodology to generate the synthetic data.

Step 1

We established all correlations between pairs of characteristics. For each pair we calculated the chi-squared statistic as a measure of correlation. For each characteristic, we then identified its two highest chi-squared statistics and selected the corresponding pairs.

We then created a graph of nodes and edges, where nodes represent characteristics and edges represent the selected pairs.

The graph represents the high correlations between characteristics in the real data that we replicate when generating the synthetic data. In the case of chi-squared statistics that are not statistically significant (close to zero) we ignore the corresponding pair. This may lead to some characteristics with one or no edges.

We then introduced directions for the edges. Each edge is directed so that it points from a node with a larger number of edges to a node with fewer edges. The result is a directed acyclic graph, i.e. a directional graph where no path starting from a node can lead back that same node.

This method is applied to each stratum of the dataset, creating a directed graph for each stratum. During synthesis, the directions in the graph determine the order in which characteristics are generated for the records in that stratum.

For example, we apply the method to the breast cancer stratum of the tumour table. The final result is a directional graph as shown in the figure below, in this graph, each node is a characteristic of the tumour table and each edge represents a highly correlated pair.

Direction has been given to edges so that nodes point to other nodes with fewer edges. It is worth noting the nodes on the far right with no edges. Some of these are rare in breast cancer, for example Gleason only applies to prostate cancer so there are no correlation found with other characteristics. Similarly, breast cancer cases would be nearly all female because male breast cancer is very rare so it also has no correlations with other characteristics.

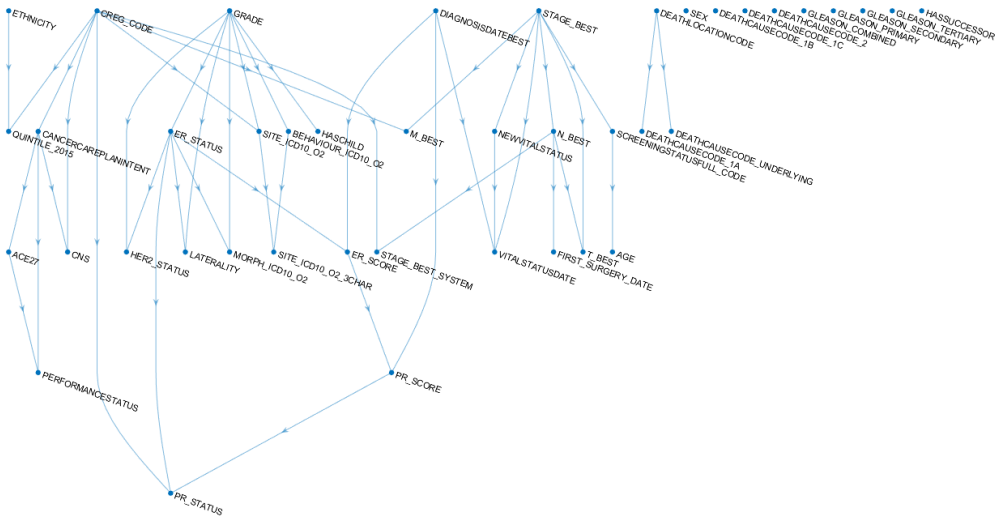

Figure 1: Directed acyclic graph for the breast cancer stratum of the tumour table.

Step 2

We then prepared a statistical model of the data to be used during data synthesis. Based on the directional graph, we calculate the conditional probability distributions of each characteristic conditional on its parent characteristics, i.e. the characteristics with edges pointing to it. For characteristics in the directional graph which have no edges pointing to them, known as the root nodes, we calculate the unconditional probability distribution.

For each subsequent characteristic, we divided the table records into groups according to distinct values of its parent characteristics. From each of these groups, we calculated the conditional probability distribution for the selected characteristic given the fixed parent value of that group. These conditional probability distributions are used during synthesis to generate values for the selected characteristic so that the correlation with its parents is preserved in the Simulacrum data.

If the group of records corresponding to a conditional probability distribution is very small, the distribution may not be statistically meaningful.

To ensure that no groups were too small, we applied a clustering algorithm within each parent characteristic. Using the clustering algorithm, we combined distinct characteristic values and gave them a single 'clustered value', so that the corresponding groups have at least 50 records. The conditional probability distributions for a given characteristic are based on the clustered values.

The merging of characteristic values for corresponding groups of records is based on the following hierarchical clustering algorithm:

consider two characteristics A and B where A points to B in the directional graph and $a_1, \ldots, a_m$ are the distinct values taken by A. Each distinct value a_i of A which has less than 50 corresponding records is merged with its next nearest value, a_j. The distance between a_i and a_j is measured as the overall difference (more specifically, the Euclidean distance) between the conditional probability distributions of B given that A is a_i and of B given that A is a_j. We continued to merge the distinct values of A with fewer than 50 corresponding patient records until all underlying groups are sufficiently large.

Taking the example of the breast cancer stratum shown in Figure 1 above, we can take a closer look at 3 nodes: the overall stage of the cancer (*STAGE_BEST*), the size of the tumour (*T_BEST*) and the degree of spread of cancer to lymph nodes (*N_BEST*). Their pairwise correlations can be represented as a subgraph of the directional graph.

This subgraph is shown in Figure 2 below, where the nodes *STAGE_BEST*, *N_BEST* and *T_BEST* are instead labelled *Stage*, *N Stage* and *T Stage*, respectively. The associated directed edges show that *Stage* is pointing to the *N Stage* and that *Stage* and *N Stage* are pointing to the *T Stage*.

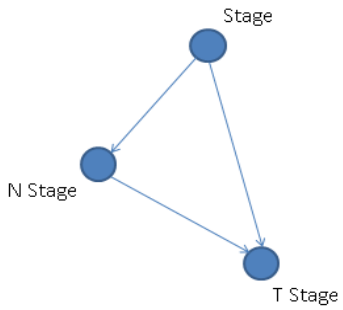

Figure 2: subgraph of the directional graph in Figure 1 representing the high correlations between Stage, N Stage and T Stage.

We then calculated the probability distributions of characteristics. Starting with *Stage*, we computed the unconditional probability distribution of *Stage*, followed by the conditional probability distribution of *N Stage* for given clustered *Stage* values. Similarly, we computed the conditional probability distribution of *T Stage* for given clustered *stage* and *N Stage* values.

Tables 1 and 2 in Appendix 1 show the clustered values of *N Stage* and *Stage* after running the clustering algorithm. These clustered values are then ready for use as parent characteristics for *T Stage.* There are several distinct *N Stage* values with fewer than 50 corresponding patient records in the real data; these are combined and given the cluster value '2b'. Similarly, distinct *N Stage* values are combined and given one of two cluster values, '1C' or '4S'. Distinct values with more than 50 corresponding records are not combined and are assigned cluster values equal to their own value. The conditional probability distribution is then computed for *T Stage* for given clustered *stage* and clustered *N Stage* values from these clustered values to be used in the generation of *T Stage.*

Step 3

The data model from Step 2 is used to generate the data. The process involves generating values of specific characteristics for each stratum of the synthetic table records. Values were generated for all the records in a stratum, one characteristic at a time, in the order in which the characteristics appear in the directional graph starting with the root nodes.

During data generation, for each characteristic and each synthetic record we selected a value from the saved probability distribution and considered other characteristic values that we conditioned on, already simulated for the given record.

> **Example:** Let us consider an example of generating a characteristic value for a virtual tumour in the breast cancer stratum of the synthetic tumour table. The characteristic we generate is *T Stage* and its parent characteristics are *stage* and *N Stage* as shown in Fig 2. The first virtual tumour has overall stage 1 and no degree of spread to the lymph nodes as already generated, i.e. clustered *Stage* value of '1' and clustered *N Stage* value of '0'. We then referred to the conditional probability distribution of *T Stage* conditional on these clustered values, calculated earlier in Step 2. This distribution is represented in Table 3 in the appendix. To generate the *T Stage* for this virtual patient we randomly sample from this distribution. This is repeated for all synthetic records in the stratum.

Summary

In these steps, we have outlined the methodology used to:

- create a single Simulacrum data table
- explained how to identify desirable statistical properties in the real patient data
- described the process of preserving these properties while generating the synthetic data

This step was repeated for each group to create the resulting synthetic dataset.

Acknowledgements

I am very grateful to all my colleagues in HDI and PHE and all those in the wider academic and industry community who have helped with this work.

Special thanks is due to our core team led by Cong Chen, with input from Paul Clarke, Sally Vernon and Jo French.

Tables

Distinct N Stage Values	Number of Patient Records with N Stage Value	Clustered N Stage Values	Distinct Value to Cluster Value Mapping
0	74195	0	0 -> 0
1	20054	1	1 -> 1
1A	1	2b	1A, 2b, 2c, N0, N1, X -> 2b
1a	6773	1a	1a -> 1a
1b	166	1b	1b -> 1b
1c	147	1c	1c -> 1c
1mi	3206	1mi	1mi -> 1mi
2	3250	2	2 -> 2
2a	2265	2a	2a -> 2a
2b	37	2b	3 -> 3
2c	3	2b	3a -> 3a
3	1490	3	3b -> 3b
3a	1283	3a	3c -> 3c
3b	85	3b	Undefined -> Undefined
3c	162	3c	
N0	1	2b	
N1	1	2b	
X	10	2b	
Undefined	15397	Undefined	

Table 1: Clustering of *N Stage* values to be used when determining the conditional probability distribution of *T Stage*. All distinct values that have fewer than 50 corresponding patient records are combined and assigned the clustered value '2b', while the rest are assigned clustered values the same as their own value.

Distinct *Stage* Values	Number of Patient Records with *Stage* Value	Clustered *Stage* Values	Distinct Value to Cluster Mapping
0	545	0	0 -> 0
1	864	1	1 -> 1
1A	46311	1A	1A -> 1A
1A1	14	1C	1A1, 1C, 4B, 4C -> 1C
1A2	4	4S	1A2, 1B1, 2A1, 2A2, 2C, 2, E, 3E, 3S, 4A, 4S -> 4S
1B	1885	1B	1B -> 1B
1B1	1	4S	2 -> 2
1C	25	1C	2A -> 2A
2	325	2	2B -> 2B
2A	31473	2A	3 -> 3
2A1	3	4S	3A -> 3A
2A2	9	4S	3B -> 3B
2B	15060	2B	3C -> 3C
2C	8	4S	4 -> 4
2E	1	4S	? -> ?
3	115	3	U -> U
3A	6524	3A	
3B	2264	3B	
3C	2255	3C	
3E	1	4S	
3S	2	4S	
4	6369	4	
4A	4	4S	
4B	9	1C	
4C	10	1C	
4S	25	4S	
?	14012	?	
U	408	U	

Table 2: Clustering of distinct Stage values to be used when determining the conditional probability distribution of T Stage. Distinct values that have fewer than 50 corresponding patient records are combined into one of two groups which are assigned clustered values '1C' or '4S', while the rest are assigned clustered values the same as their own value.

Values of *T Stage*	Probability of Choosing *T Stage* Value for a Virtual Patient (%)
1	38.8
1a	5.3
1b	15.0
1b1	0.1
1c	34.4
1mi	2.3
2	3.2
T1	0.3
Undefined	0.5

Table 3: The conditional probability distribution of T Stage given that Stage is 1 and N Stage is 0. For a virtual patient whose Stage value is 1 and N Stage value is 0, their T Stage value will be chosen this.