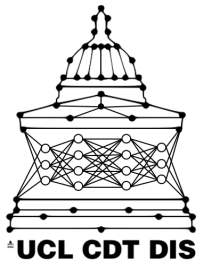

UCL CDT DIS Note

UCLCDTDIS-2019-XX

29th April 2019

valtech_

Exploration of the Simulacrum Artificial Patient-Like CancerData

Ishan Khurana[1], Jeremy Ocampo[1], Katya Richards[1], Cathal Sweeney[1], Jonathan Holdship[1], Jeremy Yates[1], and Jason Ward[2]

[1]University College London
[2]Valtech

Publicly available Exploratory Data Analysis tools have been developed for the Simulacrum synthetic cancer dataset. This will serve as a base upon which future collaborators may build. Patient pathways were constructed from the Simulacrum dataset and a method for clustering the pathways was created using an LSTM autoencoder, this allows for segmenting pathways with differing lengths, ordering and combination of events that constitute a pathway, in addition this method has opened up a number of possible ideas for generation of pathways that can be developed in the future.

Contents

1 Introduction

When considered separately for males and females, of the top ten most common causes of death in the UK four of them are cancers [1]. Cancer accounts for 30.0 % of deaths in males and 24.8 % of deaths in females in 2015 [1]. For this reason it is of great interest to society as a whole to understand as much as we can about the journey of a cancer patient. However for this very same reason there is a large amount of cancer data. Hence it would be potentially quite beneficial to grant data scientists access to cancer data.

Simulacrum is a synthetic cancer dataset developed by Public Health England (PHE) [2], which became publicly available for the first time in late 2018. The Simulacrum data is in the same format as real patient data held in the National Cancer Registration and Analysis Service (NCRAS), but as it is simulated data it is not bound by patient confidentiality constraints. In this way it opens up cancer data to data scientists without the need to get clearance to access real patient data. One of the goals of the project is to allow data scientists to develop models that can be submitted to run on real cancer data.

Our aim is to create publicly available tools to help anyone using the Simulacrum dataset in the future. We also perform some exploratory artificial intelligence methods. Finally, we have engaged healthcare professionals to get a better understanding of the limitations of the data.

Our work was written in Jupyter Notebooks and can be found on GitHub. [3]

2 The Dataset

The data is simulated in such a way as to mimic statistical distributions of real data. That said, for very complex inference (e.g. deep learning) it is possible that there may be discrepancies between real and simulated data. Simulacrum have released an informal white paper [4] giving a description of how the data was generated.

2.1 Structure

The dataset is comprised of seven tables which can be linked together with linking tables as shown in Fig. 2.1. This is the exact same table structure as NCRAS data is stored in, with the only difference being the prefix "sim_" to emphasise that this is simulated data. Each table contains a specific type of information e.g SIM_AV_PATIENT records each patient's sex, ethnicity, death cause (if applicable).

2.2 Terminology

- SACT: Systematic Anti-Cancer Therapy i.e. chemotherapy drugs
- Regimen: A prescribed course of medical treatment which may contain numerous cycles and a combination of different drugs

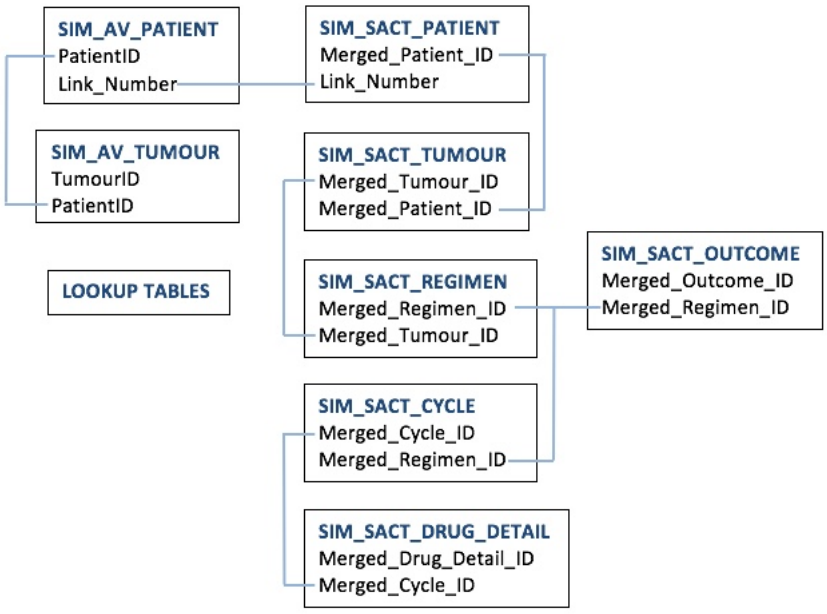

Figure 2.1: Tables within the Simulacrum dataset, showing how they can be linked together. Figure reproduced from Simulacrum [2]

- Cycle: A period of (chemotherapy) treatment. A collection of cycles, with rest periods in between, makes up a regimen.
- Pathway: The collection of all events associated with a patient
- Event: A medical event classified by the SACT tables in Figure 2.1, some examples are: diagnosed with breast cancer, start of trastuzumab regimen, drug of cyclophosphamide taken.

2.3 Challenges

There are some challenging aspects of this dataset. We will list some here:

- Misspelling of drug names
- Nonsense data e.g. some patients who are listed as "Alive" have an associated death cause
- Lots of NaNs (incomplete data)
- It is difficult to tell if the issues listed above are the result of the simulation mimicking real cancer data or if these issues exist solely in the simulated data

2.4 Distributions

In order to have a feel for what is in the datasets and what parts of the data are most robust we decided to plot a bunch of bar plots and compare them with real life data. Figure 2.2 shows the simulated distribution of cancers by age, this can be compared with the real life distributions from cancerresearchuk [5][6], the distributions compare well except cervical cancer. The difference is that the count for the peak of cervical cancer is $\sim 1\%$ of the max frequency from cancerresearchuk compared to $\sim 30\%$ from the simulation in Figure 2.2.

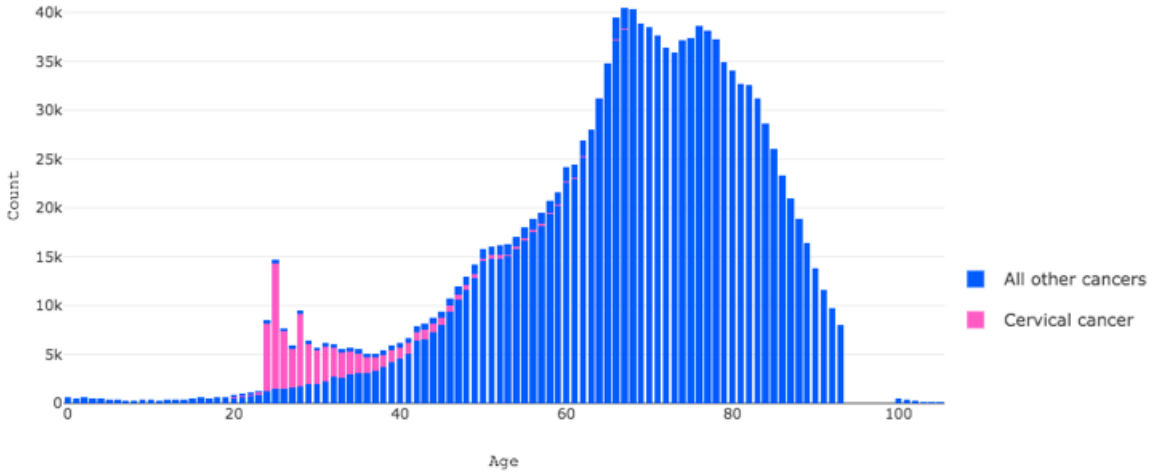

Figure 2.2: Stacked bar plot of frequency of diagnosed cancers by age

Figure 2.3 shows that the distribution of deathcauses agree well with cancerresearchuk [7], except that the simulation has simulated females with prostate cancer and males with ovary cancer. We have confirmed with PHE that that these incidences do exist in the real dataset and the reason they do not appear in [7] is that they are excluded from incidence reports.

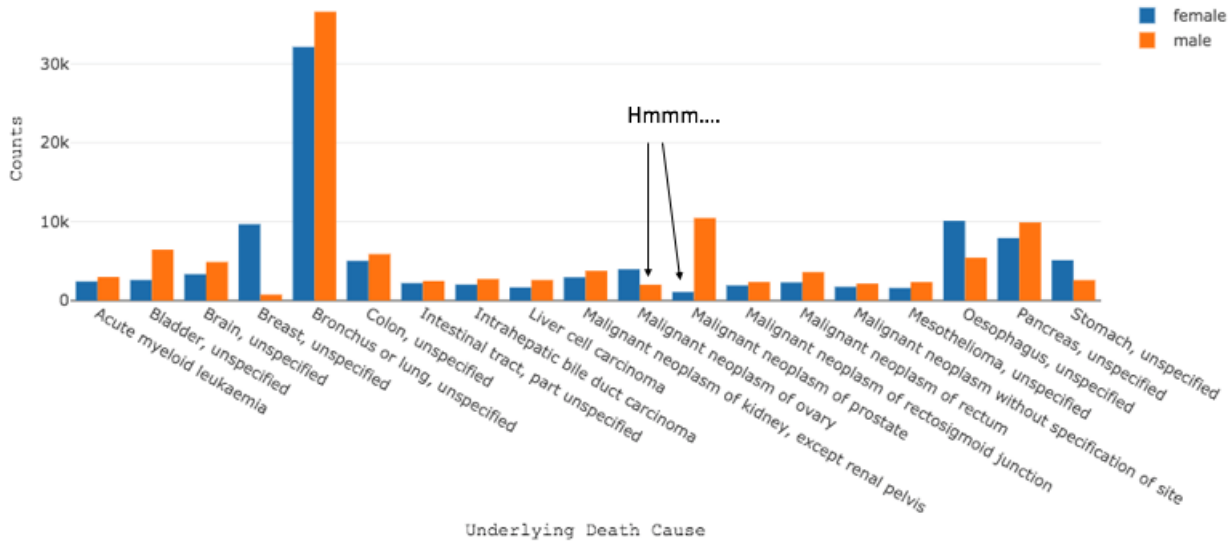

Figure 2.3: Frequency of death by underlying death cause

3 Events and Patient Pathways

3.1 Constructing Pathways

A patient pathway is the route that a patient will take in the course of their treatment, this route is made up of the events in the patients history. We were able to reconstruct the sequence of events in the pathways of each patient by linking the SACT tables in Figure 2.1. Some example pathways are shown in Table 1

PATIENT 1	PATIENT 2	PATIENT 3
diagnosis C50		
regimen TRASTUZUMAB	diagnosis C50	regimen FEC
drug STEROID	regimen FEC	drug TRASTUZUMAB
drug BORTEZOMIB	drug TRASTUZUMAB	drug CYCLOPHOSPHAMIDE
drug NOTCHEMO	drug NOTCHEMO	drug NOTCHEMO
drug FLUOROURACIL	drug STEROID	diagnosis C50
drug CYCLOPHOSPHAMIDE	drug FLUOROURACIL	regimen TRASTUZUMAB
drug EPIRUBICIN	drug NOTCHEMO	drug EPIRUBICIN
regimen TRASTUZUMAB	regimen dose reduction Y	drug CYCLOPHOSPHAMIDE
drug FLUOROURACIL	drug NOTMATCHED	drug NOTCHEMO
drug CYCLOPHOSPHAMIDE	drug DOCETAXEL	drug NOTCHEMO
drug DOCETAXEL	drug STEROID	drug NOTCHEMO
drug NOTCHEMO	drug CYCLOPHOSPHAMIDE	drug DOCETAXEL
drug DENOSUMAB	regimen FEC	drug CYCLOPHOSPHAMIDE
drug DOCETAXEL	drug TRASTUZUMAB	drug NOTCHEMO
drug EPIRUBICIN	regimen outcome 0	drug NOTCHEMO
regimen time delay Y		drug FULVESTRANT

Table 1:

Three example patient pathways shown including drug names and diagnoses. Some specialist vocabulary:
diagnosis C50 - This is the ICD10 code [8] for malignant neoplasm of breast
regimen FEC - FEC are the initials of the drugs used in this regimen
regimen outcome 0 - An outcome of 0 means the regimen was completed as prescribed
Y - Yes

As these pathways are just sequences of events in chronological order, we can learn about their structure by applying sequence models to them. The types of events we included from the dataset are diagnosis, start of a regimen, drug administered, dose reduction, delaying the time between administration dates and the outcome of the regimen, we have chosen these types as these are the main events in the dataset that constitute a pathway.

3.2 Giving events context

As pathways have a similar structure to sentences, we can use existing natural language processing(NLP) algorithms on the pathways by using sentences as pathways and words as events.

Word2vec [9] is a model widely used for NLP, it can be used to learn the features of words in the form of a vector such that words that frequently occur in the same sentences (e.g. “quantum” and “physics”) and frequently appear within some number of words from each other, have vectors that are close to each other. It consists of a 2 layer neural net in which the input is a word from a sentence and the output is another word from the same sentence, in this way the neural net “learns” the context of each word i.e. which group of words appear frequently in the same sentences. After the neural net is trained on all sentences, you can extract the vector of features(i.e. word embedding) for each word from the hidden layer of the neural net.

Using Word2vec on pathways, we can give each unique event a vector of features that is extracted from the Word2vec neural net such that groups of events will have similar vectors if they frequently appear together in the same pathway and if they are frequently within some number of events from each other in that pathway. In Figure 3.1, the vectors of events have been plotted, and each region of the space corresponds to groups of events that frequently occur together in a pathway. For example there is a region with breast cancer events such as the administration of the drugs fulvestrant and eribulin, these drugs are commonly used in breast cancer. These points are 100 dimensional vectors that are extracted from the Word2vec neural net where each dimension represents a feature of that event, for example a feature that might have been learned by the neural net is if that event occurs frequently with breast cancer or not, note that these features are not manually chosen but rather learned by the neural net.

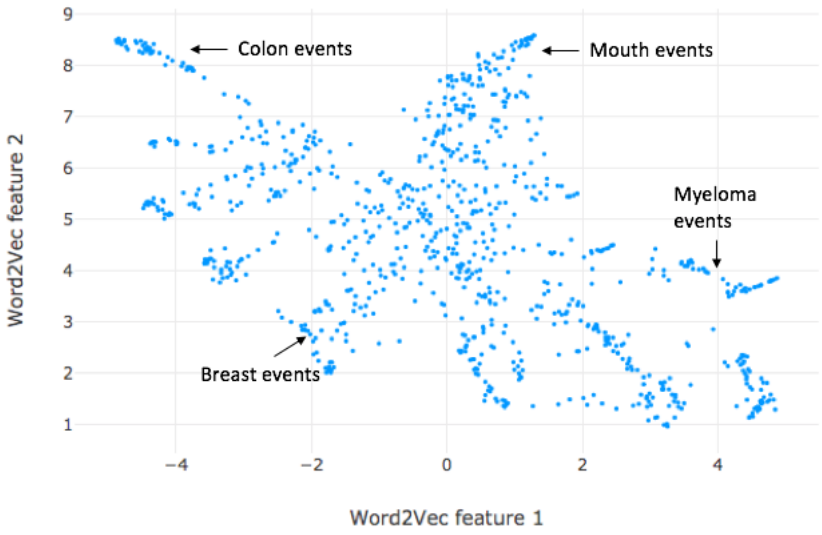

Figure 3.1: Each point here represents a unique event e.g diagnosis of breast cancer, administration of the drug fulvestrant, start of the regimen trastuzumab. The space that they are in is 100 dimensional and is learned by the Word2vec neural net. This 100D space is reduced to 2D by Umap [10](explained in appendix A.1) and is plotted here.

3.3 Pathway Visualization

To visualize the pathways we can add up all the events(their Word2Vec vectors) in a given pathway and connect the points as shown in Figure 3.2, where each line corresponds to a pathway. This shows that the pathways cluster well according to cancer as you would expect each cancer to contain different drugs/regimens; this information is not really informative so we decided it would be more useful to work on a single cancer and investigate the structure of the pathways corresponding to that cancer. We chose to work on breast cancer and clustered it's pathways.

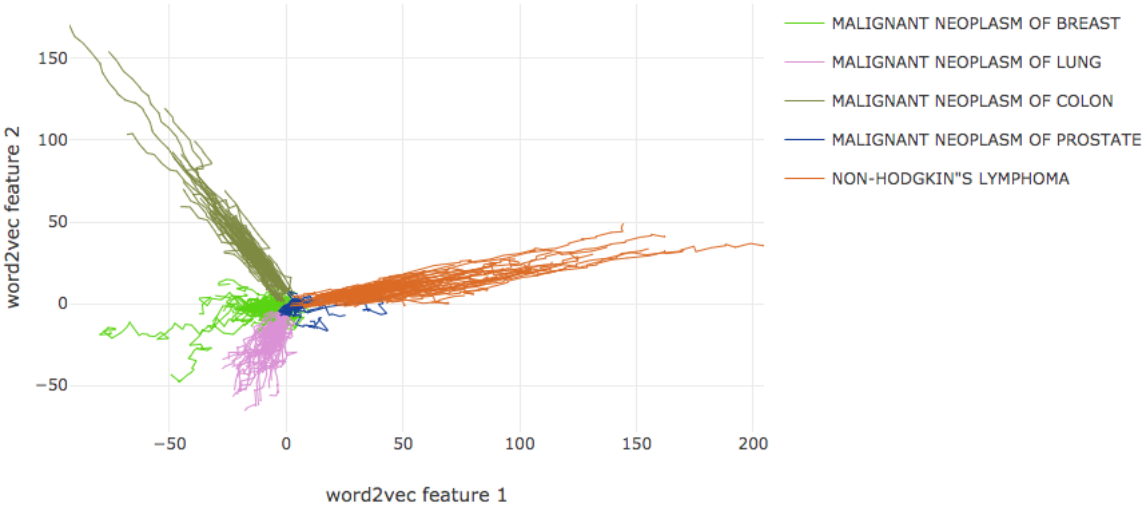

Figure 3.2: Each line here is an individual pathway of a patient, and that pathway is constructed by adding up all the event vectors from Figure 3.1 that are in that patients pathway. This is colour coded by 5 different cancers shown in the legend.

4 Pathway Clustering

4.1 Clustering Methods

We wanted to see if the pathways can be clustered or segmented based on their features, this would give us an overview of the different types of pathways(if any) and how distinct they are. In addition, finding a method to obtain the space of features of patient pathways leads to ideas of generating pathways which could lead to a variety of uses. In order to compare pathways we used a model such that for a given pathway of any length, the model can produce a 150 dimensional vector of features for that pathway. This was done using an autoencoder [11] which is a neural net designed for unsupervised learning.

In order for the autoencoder to work with sequences we used a recurrent neural net(RNN) [11], these are widely used in NLP and can be thought of as a feed forward neural net "unrolled across time" because you are inputting words to the neural net as time moves forward, this is what we want as the structure of the pathways are the events of a patient across time. The structure of the autoencoder used is shown in Figure 4.1. We want to train it to output the same sequence of events as the input sequence. In this way all the information from the input pathway will be encoded in a 150 dimensional vector - the encoding. The encoding should then have information on the input sequence which can be used by the decoder(top layer of Figure 4.1) to reconstruct the input sequence.

The RNN we used is composed of LSTM cells [11][12]. Why LSTM? Because an LSTM cell is designed such that it can handle long/short sequences in a way that the cell "remembers" events that has happened in the far past. For example if their are a lot of sequences of the form

```
[ diagnosis C50, regimen Trastuzumab, .......  , drug Docetaxel ]
```

then the neural net is trained such that the initial information - diagnosis C50, regimen Trastuzumab is retained until it reaches drug Docetaxel, in this way the neural net "learns" that these initial two events are correlated to the final event.

After training we can extract the vector of features(the encoding) and cluster them via a k-means [13] clustering algorithm to see what sequences are similar and also investigate the structure of the sequences/pathways.

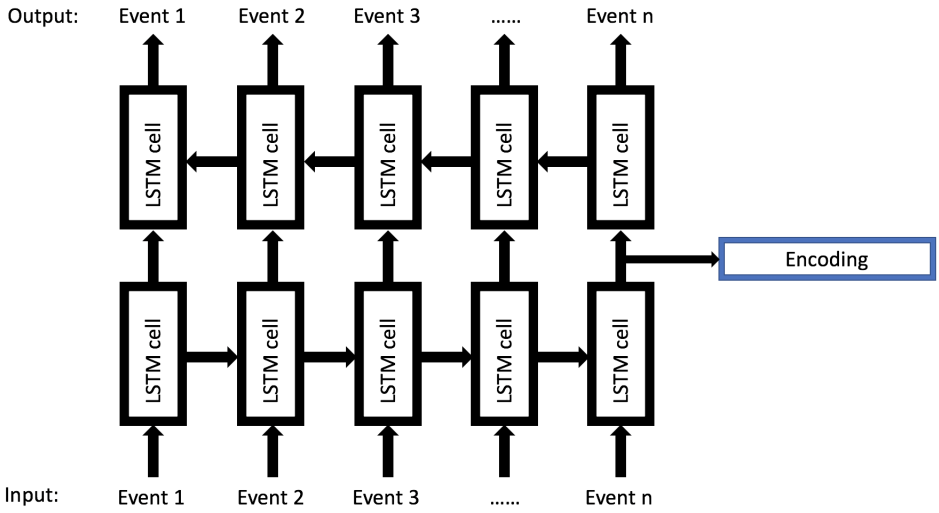

Figure 4.1: This is the structure of the LSTM autoencoder model that was trained for unsupervised learning. The encoding is what we wanted to extract and to use as the features of a sequence. The input is the sequence of Word2vec vectors of the events in a pathway. This model is sequence length independent. Note that the same weights are being used and trained in the LSTM cell.

4.2 Clustering Results

We were able to show that the pathways of breast cancer patients cluster well (Figure 4.2), and that each cluster corresponds to a unique structure of a pathway. Some examples of what different clusters represent are:

- pathways with an outcome of having an acute toxicity to the chemotherapy they've undergone
- pathways which have treatment prior to diagnosis
- pathways which have taken a particular combination of drugs

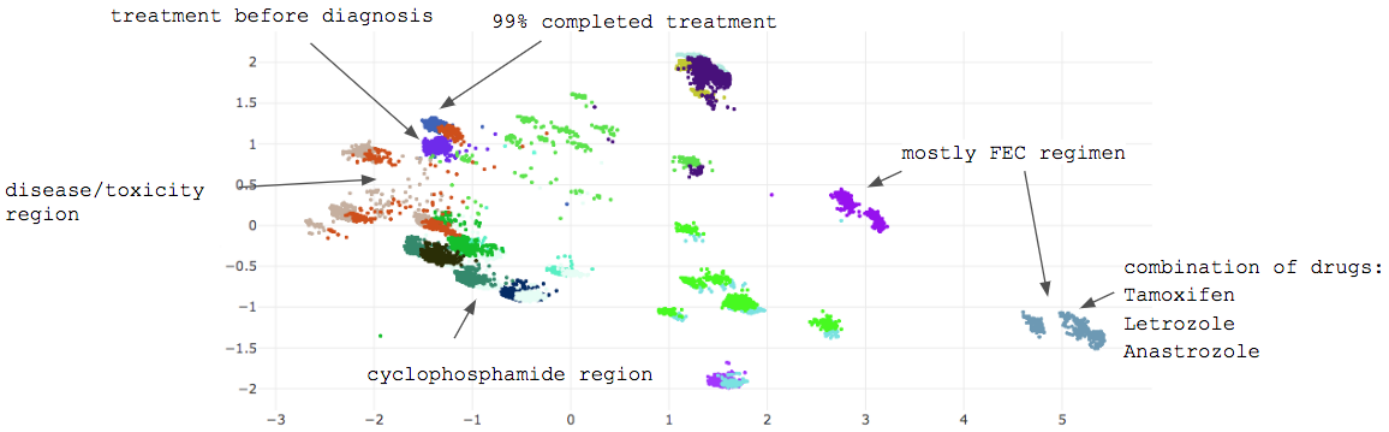

Figure 4.2: Each point here represents a sequence/pathway of a patient, it is evident that the points are not randomly spaced, instead they are clustered in different regions and the different colours here represents the different clusters. The space that they are in is a 150 dimensional space of features of the sequences e.g. one of the dimensions might be how long the sequence is, these are learned by the RNN autoencoder. The 150 dimensions has been reduced by PCA [14] into 2 dimensions as shown in the figure. A different dimension reduction technique is shown in appendix A.2.

Finding out what each cluster means is difficult as we have to manually search the features that distinguishes them. Each cluster mostly have similar distributions of the types of drugs/regimens they have undertaken which should not be a surprise as the distributions of drugs/regimens should be similar because all patients here have a diagnosis of breast cancer, it is the order of events in the pathway that distinguishes the clusters. For example one cluster has the sequence of events

```
[ diagnosis C50, regimen FEC, drug Trastuzumab ]
```

inside most of the pathways in that cluster(in exactly that order). We also found that most clusters have pathways that are 300 days or 700 days long.

The model we have used is very simple and has not been fine-tuned, this is because the simulacrum data is simulated and is in it's early stages, if we fine-tuned our model there is a good chance it will over-fit to the simulation and not the real data, and also not fine-tuning means it is less likely that the model will over-fit to the dirtiness of the data.

5 Regimen Outcomes

An investigation of the treatment completion rate has been carried out using the simulated data set. Six cancers were studied and the results are discussed in this section.

Figure 5.1 below shows the percentage of patients that successfully complete a SACT. The patients that do not complete a SACT can have one of the following outcomes: (1) patient dies, (2) disease progresses during chemotherapy, (3) acute chemotherapy toxicity, (4) technical or organisational problems, (5) patient stopped or interrupted treatment. These were grouped together in this preliminary analysis since the data set did not have enough entries for the individual failed outcomes to give a statistically significant result.

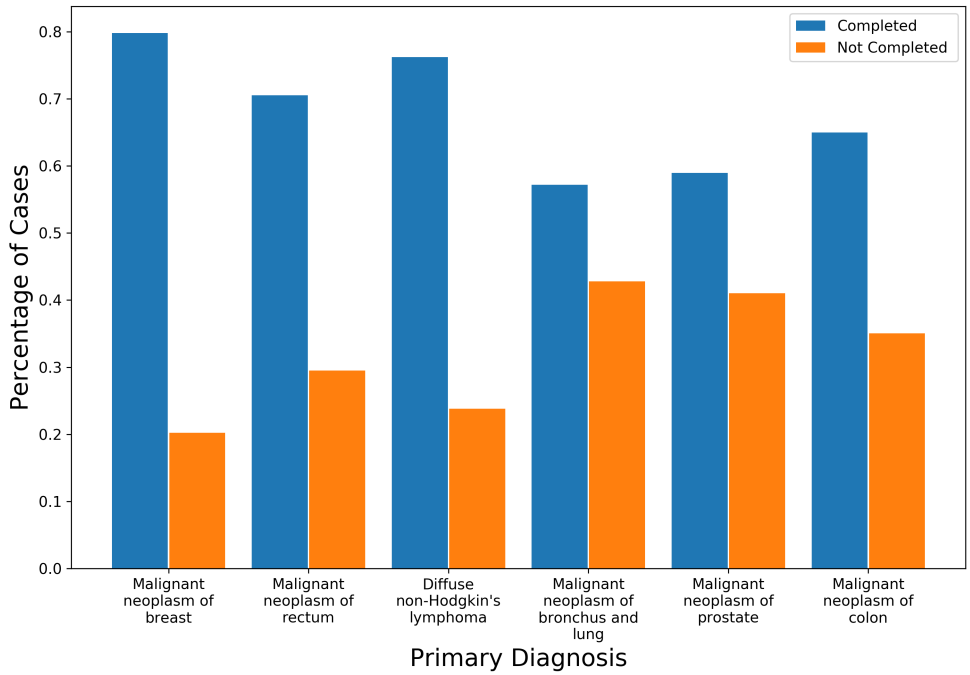

Figure 5.1: The Systematic Anti Cancer Treatment completion rate for the six most common cancers in the simulacrum data set.

The data shows that Malignant Neoplasms of Bronchus and Lungs have the lowest completion rate in the cancers studied. To find a possible cause for the differences in the completion rates, the completion rates of treatments using a particular chemotherapy drug were studied. These completion rates have been plotted in figure 5.2

The data generally shows that the type of drug available has an effect on the likelihood of completing a treatment. Drugs used to treat cancers with lower treatment completion rates, appear to have low overall completion rates when applied to other cancers as well. This effect should be studied further and could provide predictive power in a classification algorithm.

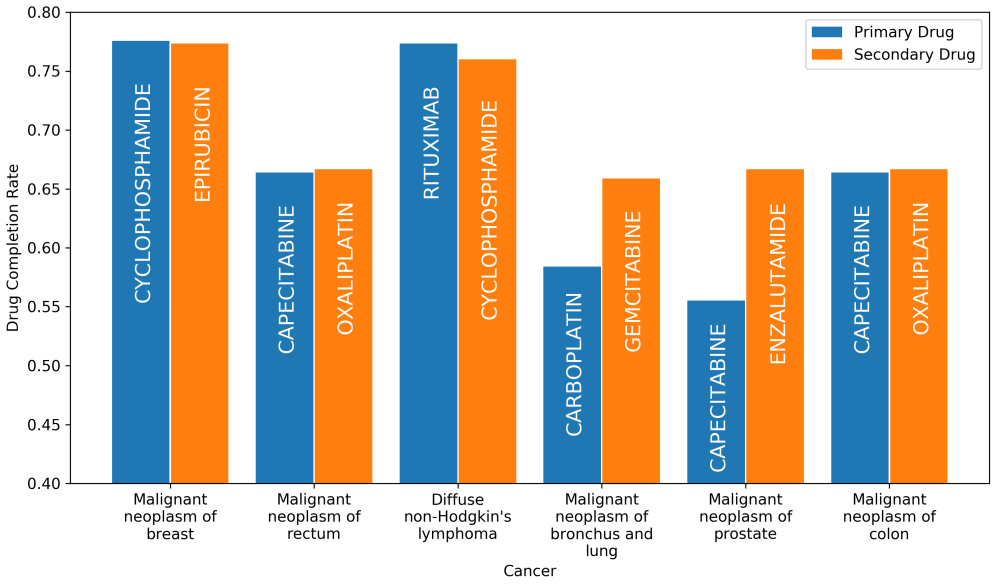

Figure 5.2: The completion rate for the primary and secondary chemotherapy drugs used to treat the cancers studied. Drug names are shown in the bars. Drug completion rates are calculated using the outcomes from prescription for any type of tumour and not just the ones studied.

6 Conclusions

We have explored and cleaned some parts of the datasets and have found that a lot of the distributions of the simulation compare well with real data although there are some parts of the simulation which did not make sense, this could be either from the simulation being wrong or the simulation has simulated dirty data correctly, or both.

We have implemented a basic LSTM autoencoder model for clustering the pathways of breast cancer and the clusters appear to be distinct in their lengths,combination and ordering of events. An immediate use of the clusters is to look at outliers of a cluster and seeing what was different in their pathway and finding the cause of that difference, note that some outliers may not be entirely meaningful for example a pathway with only one event would be an outlier. Their are a lot of possibilities for future work and uses, if the simulation and datasets can become more accurate then the models that are made can be more fine tuned and developed further, what we have done in this project are some basic implementations/analysis as this is just the beginning of finding out what can be extracted from this dataset. Some possibilities for future work could be:

- Given some initial conditions of a patient, can we tell what their pathway would look like with some probability, using the clusters and supervised learning?
- Can we predict the exact sequence that will happen to a patient?

- Can we generate pathways that have the best outcomes? for example this could be done by iterating through the space of good outcomes in figure 4.2, and then obtaining the pathways from that space.
- Can we generate pathways that lead to toxicity/death that are not obvious, in this way preventing that pathway?
- Can we generate pathways that are short and efficient (less drugs) but also have good outcomes?

7 Acknowledgments

Data for this study/project/report used artificial data from the Simulacrum, a synthetic dataset developed by Health Data Insight CiC derived from anonymous cancer data provided by the National Cancer Registration and Analysis Service, which is part of Public Health England.

Thank you to Cong Chen of Public Health England for helping us navigate Simulacrum and answering our questions.

Also to Andre Vauvelle who shared his work and insights from being a Machine Learning Research Intern at Public Health England.

We would like to extend gratitude to Katie Tucker, Graham Roberts and Emma Bointon (Guy's and St Thomas' NHS Foundation Trust) for kindly meeting with us to give us an insight into how cancer data is recorded and stored. This provided us with a better appreciation of how cancer data is recorded in the "real world".

Finally to Jason Ward and Jon Holdship for their guidance. Thanks to them we have learned new data science skills and tools which will prove invaluable to us in future endeavours.

A Dimension reduction

A.1 Methods

Different methods of dimension reduction were used in this project, brief explanations of the methodology behind them are given below:

- PCA[14]: An optimal projection of the data points from 100D to a 2D plane is found such that the points are maximally separated. This is what you want for visualizing the clusters because if you projected the data points to a plane such that the points are not separated well, you might project onto a plane such that the clusters bunch into some region which makes it hard to tell if those clusters are separated in the visualization.
- UMAP[10] / t-SNE[15]: An optimal mapping of the data points from 100D to 2D is found such that the proportion of distances between points are conserved i.e. the distribution of points are conserved. This allows you to visualize the "shapes" of the data points in 2D. Some differences are that UMAP has a faster run time and better visualization quality.

UMAP and t-SNE are advantageous if you want to visualize the distributions and topology of the clusters in a lower dimensional space, but if you are just looking for clusters and dont care about their shapes, PCA has a faster computation time.

A.2 t-SNE for pathways

The dimension reduction of the feature vectors of the pathways gives more interesting results when done via t-SNE, see Figure A.1. The clusters seem to reveal branches instead of cluster bunches as in Figure 4.2. This is actually an interactive plot which allowed us to instantly see what the pathway of a point is by hovering over it(see patientpathways.ipynb on our github [3]). We saw that pathways which are on the same tree but on different branches have a similar first half sequence but a different second half sequence, or the other way round.

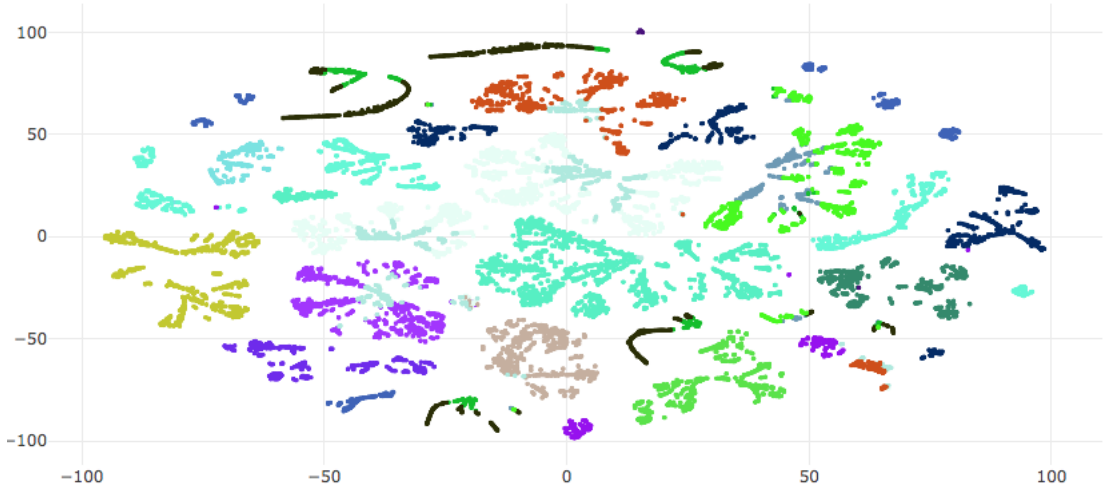

Figure A.1: Each point here represents a patient pathway. This is essentially the same plot as Figure 4.2, the difference is that the dimension reduction is done by t-SNE.

In Figure 4.1 the input/output to the autoencoder is a sequence of events in the form of their Word2vec vectors, on those vectors we have attached the number of days that has passed for that event to happen, we believe that this is the cause of the branches appearing just from the fact that the branches do not appear if you do not attach the number of days onto the vectors. Points that are close to each other are similar pathways and so we hypothesize that as you are moving through a branch you are going through the number of days, this is because a difference in nearby points are the number of days in that sequence. More analysis is needed for a better interpretation of Figure A.1.

REFERENCES

[1] UK Government. *Major causes of death and how they have changed*. URL: https://www.gov.uk/government/publications/health-profile-for-england/chapter-2-major-causes-of-death-and-how-they-have-changed.

[2] Simulacrum. *Background on the Simulacrum*. URL: https://simulacrum.healthdatainsight.org.uk/background-on-the-simulacrum/.

[3] *Git repository for the work in this report*. URL: https://github.com/UCL-simulacrum/EDA.

[4] L Frayling. *Generating the Simulacrum A methodology overview*. Nov. 2018. URL: https://simulacrum.healthdatainsight.org.uk/wp/wp-content/uploads/2018/11/Methodology-Overview-Nov18.pdf.

[5] *All cancers incidence by age*. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/age#heading-Zero.

[6] *Cervical cancer incidence by age*. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/cervical-cancer/incidence#heading-One.

[7] *Common causes of cancer death*. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-cancers-compared#heading-Zero.

[8] *International Classification of Diseases (ICD) Information Sheet*. URL: https://www.who.int/classifications/icd/factsheet/en/.

[9] *Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781.*

[10] Leland McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[11] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[12] Sepp Hochreiter and Jurgen Schmidhuber. *Long Short-term Memory*. Dec. 1997. DOI: 10.1162/neco.1997.9.8.1735.

[13] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967, pp. 281–297. URL: https://projecteuclid.org/euclid.bsmsp/1200512992.

[14] Karl Pearson F.R.S. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720. eprint: https://doi.org/10.1080/14786440109462720. URL: https://doi.org/10.1080/14786440109462720.

[15] Laurens van der Maaten and Geoffrey Hinton. "Viualizing data using t-SNE". In: *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605.