# Colon Cancer Survival

*Predicting the probability of survival at two years after colon cancer diagnosis*

Elsa García Hernández

# CONTENTS

# Introduction

Colon cancer is a common and lethal disease. According to GLOBOCAN 2018 data, colon cancer is the fourth most incident cancer and the fifth most deadly cancer worldwide[1].

Colon cancer usually begins with the non-cancerous proliferation of mucosal epithelial cells. These growths are known as polyps and can grow gradually for 10–20 years before becoming cancerous. The most common form is an adenoma or polyp that originated from granular cells, whose function is to produce the mucus that lines the large intestine. Only about 10% of all adenomas progress to invasive cancer, although the risk of cancer increases as the polyp grows larger. The tumours that grow into the wall of the colon can penetrate blood or lymphatic vessels, allowing metastasis to distant organs via the blood or to nearby lymph nodes. The extent of invasion determines the staging, and thus the prognosis, of a colon cancer diagnosis.

Survival analyses are crucial to understand the situation of a disease in a population, as well as all the information about new cases and mortality. In addition, survival analyses allow us to quantify the effectiveness of the screening and treatment services. As well as other health indicators, survival analyses show variations across populations and time, and the analysis of these variations provides us with information that allows setting improvement goals in the diagnosis and treatment of a disease. One of the most popular regression techniques for survival analysis is the Cox proportional hazards regression analysis, which assesses simultaneously the effect of several risk factors on survival time.

Accurate prediction of the probability of survival in patients with cancer remains a challenge due to the ever-increasing heterogeneity and complexity of cancer, treatment options and patient populations. However, the new technologies that are emerging could contribute to address these issues. Healthcare is a field that is potentially suitable for many applications of artificial intelligence tools and techniques, since the quantity of healthcare data that is collected has remarkably increased in the recent years. Thus, artificial intelligence tools could play a key role in using this information to improve our quality of life and enhance the quality of decision-making in the healthcare system.

Machine learning is a subset of artificial intelligence that offers a wide range of potential applications in healthcare. Machine learning models could be useful to conduct survival analyses, thus, the aim of this project was to predict the probability of survival at two years after colon cancer diagnosis by training a machine learning model.

[1] Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Prz Gastroenterol*. 2019;14(2):89-103. doi:10.5114/pg.2018.81072

# Simulacrum Database

This project has been developed using a database called Simulacrum. The Simulacrum database imitates some of the data held securely by the National Cancer Registration and Analysis Service (NCRAS) by the National Disease Registration Service, which is part of Public Health England (PHE). NCRAS collects data on all cancers diagnosed in England. This is then linked with other data from the United Kingdom National Health Service (NHS) to create a large and complex database that is held in the Cancer Analysis System (CAS). The Simulacrum has been designed to mimic some of the data held on the CAS.

The data in the Simulacrum is entirely artificial. It does not contain data about real patients, so users can never identify a real person. It is free to use and allows anyone who wants to use record-level cancer data to do so, safe in the knowledge that while the data feels like the real thing, there is no danger of breaching patient confidentiality. The shape of the data is the same as the real one so that it can be used to run analyses that (with the right permissions and ethical approval) could be run on the real data.

The Simulacrum was developed by Health Data Insight CiC (HDI), a Cambridge-based social enterprise, with support from AstraZeneca and IQVIA. Further information about the database can be found here: https://simulacrum.healthdatainsight.org.uk/

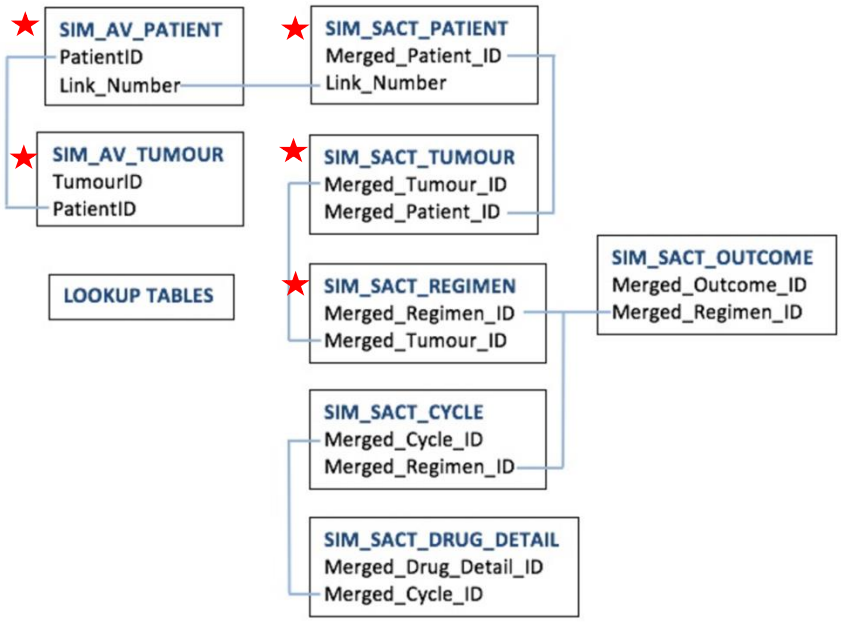
# Raw Data Description

The Simulacrum imitates data about tumours diagnosed in England in 2013, 2014 and 2015. The vital status of each synthetic patient up to the end of 2017 has also been simulated, so survival analyses can be conducted using the Simulacrum data. There is also data about treatments with chemotherapy, known as Systemic Anti-Cancer Therapy (SACT), which contains information about the types and number of treatments received. SACT data has been simulated up until June 2016 so it is possible to analyse the treatments following diagnosis.

There are three sets of linked tables within the Simulacrum database:

- Cancer registration tables (SIM_AV)
- SACT tables (SIM_SACT)
- Lookup tables

The tables are linked as follows:




For simplicity, only the tables that are marked with a red star have been used for this project.

The SIM_AV_ tables represent the patient and tumour registration data, and the SIM_SACT_ tables represent the SACT data. These two sets of tables are linked via the patient tables by the Link_Number, which is the common identifier in both datasets.

The SIM_AV_TUMOUR table contains most of the detailed information regarding staging and pathology of the tumour. Each patient can have multiple tumours in both the registration and the SACT datasets. Aside from the linkage by the Link_Number, both the SACT and AV tables are independent, so the tumour IDs in the SIM_AV_TUMOUR table cannot be matched to the tumour IDs in the SIM_SACT_TUMOUR table. The SACT tables contain details of treatments received for patients and can be used to identify patients who were treated with certain therapeutic regimens.

The descriptions of each of the tables that have been used in this project can be found below:

- **SIM_AV_PATIENT:** Includes patient demographics, vital status date and cause of death. Contains Link_Number which allows linkage to SIM_SACT_PATIENT.
- **SIM_AV_TUMOUR:** Main tumour table containing the details of each tumour registered. Each patient may have more than one tumour. Includes tumour ID, age at diagnosis, screening status, demographics, tumour site, staging, TNM staging, grade, tumour morphology, performance status, hormone receptor status, surgery date and deprivation index.
- **SIM_SACT_PATIENT:** Contains just patient ID and Link_Number which allows linkage to SIM_AV_PATIENT.
- **SIM_SACT_TUMOUR:** Contains patient ID, tumour ID (not the same as in SIM_AV_TUMOUR table), primary diagnosis site, morphology code and consultant specialty code.
- **SIM_SACT_REGIMEN:** Contains regimen ID, benchmark group, mapped regimen, chemo radiation, intent of treatment, date of decision to treat, height and weight at start of regimen, clinical trial enrolment and start date of regimen.

# Methodology

The project was conducted using Jupyter Notebooks and Python (version 3.7.4) in a virtual environment that was accessed through Ubuntu (version 18.04.3). The stages that were followed to develop the project were the following:

**Data acquisition:** The raw data that was used for the project was downloaded from this webpage: https://simulacrum.healthdatainsight.org.uk/requesting-data/. The data was saved locally and then uploaded to a virtual environment using Ubuntu. All the analyses were conducted in this virtual environment by using Jupyter Notebooks.

**Data preparation and cleaning:** During this stage, the tables of interest were linked and the data was prepared and cleaned. Variables that contained a very high amount of missing data were removed, as well as those variables that were not potentially useful (based on my knowledge of this area).

Since each patient might have several tumours and each patient might receive several treatments for each tumour, the number of rows grew exponentially while the tables were being merged. As a result, the information of each patient ended up being displayed in several rows (depending on the number of tumours and the number of treatments). Therefore, one of the main steps of this stage was to aggregate the data at the patient level, so that in the final dataset all the information of one patient was contained in a single row. In order to achieve it, new variables were created based on the ease of creating the variable and the potential usefulness of that variable (based on my knowledge of this area).

Values that were not plausible were removed and replaced by the median values (e.g., negative values for height or weight). Missing values were also replaced by the median.

Each variable was plotted to check the distribution of the variable. In addition, the variables were plotted against the outcome (alive/dead) to have an overview of the differences by vital status for each variable.

These steps were executed in the notebooks 1 to 6.

**Analysis and model preparation:** The aim of the project was to predict the probability of survival at 2 years after colon cancer diagnosis. Thus, the outcome was a variable that indicated if the patient was alive or dead at 2 years. Patients who had died within the first 2 years and patients who had been alive for at least 2 years were included in the analysis. Collinearity issues were assessed using the variance inflation factor, and numerical variables were normalised.

A logistic regression classifier was built in first place. Taking this model as a basis, the final variables that were included in the final dataset were selected following a stepwise selection approach. Other models that were also evaluated included a decision tree, a random forest and a light gradient boosting machine classifier. The Python library that was used was Scikit-learn.

To account for the class imbalance (68.48% of patients were alive and 31.52% were dead at 2 years after diagnosis), the option class_weight='balanced' was used in all models so that the algorithm automatically adjusted weights inversely proportionally to class frequencies. The model performance was evaluated using the metrics "recall" and "F1-score" mainly. Recall is appropriate for imbalanced datasets, and the F1-score is useful since a balance between recall

and precision was important for my project. In addition, the confusion matrix and the precision-recall curve were plotted. The average precision score was also obtained from the precision-recall curve.

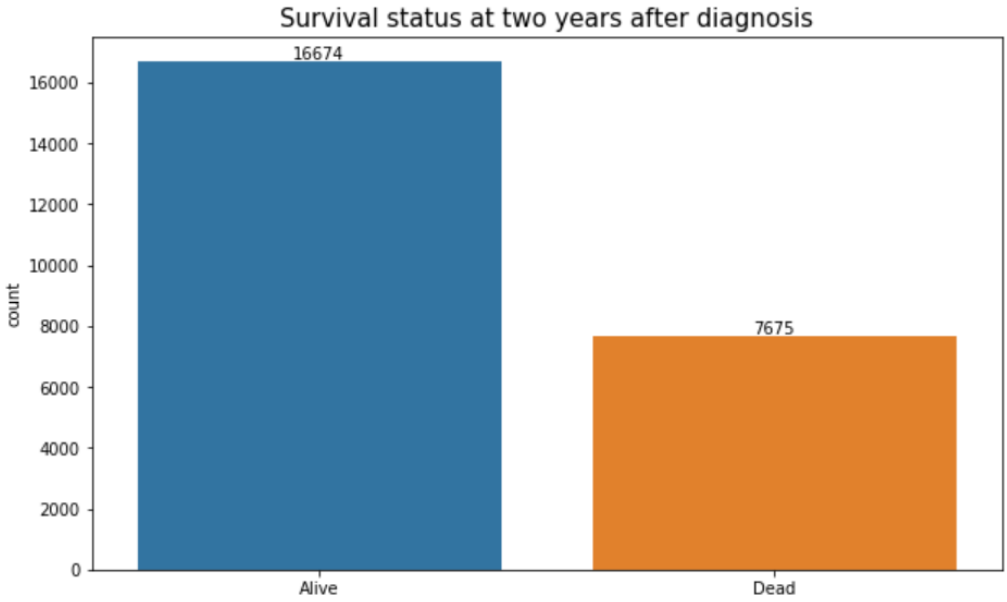
These steps were executed in the notebook 7.

**Visualisation:** To conclude the project, an interactive visualisation was created using the library Streamlit. This interactive visualisation displays a plot that shows the probability of survival at 2 years after colon cancer diagnosis, based on the inputs that the user enters in the interface.

This step was executed in the notebook 8 and the *visualisation.py* file.

## Summary of Main Results

As mentioned above, the aim of the project was to predict the probability of survival at 2 years after colon cancer diagnosis, and the outcome was a variable that indicated if the patient was alive or dead at 2 years. The following plot represents the distribution of the outcome:




As we can see above, 16674 (68.48%) patients were alive and 7675 (31.52%) patients were dead at 2 years after colon cancer diagnosis. The analysis was conducted on these patients.

The variables that were retained in the final analysis following the stepwise selection approach were the following:

- Age
- Gender
- BMI
- Number of tumours
- Time from tumour diagnosis to surgery
- Tumour stage
- Tumour grade

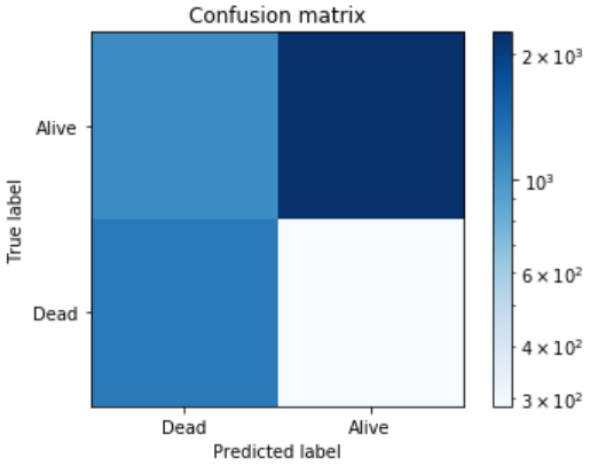
Age, gender, BMI and number of tumours were not found to be associated with the probability of survival during the stepwise feature selection approach. However, they were retained in the model because age, gender and BMI are common confounders which are typically included in the models, and the number of tumours was useful to develop the visualisation. The variables that were found to have a higher impact on the probability of survival were time from tumour diagnosis to surgery, tumour stage and tumour grade.

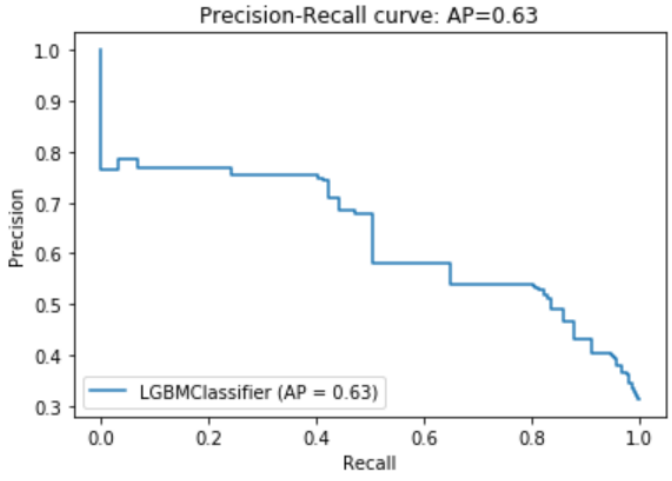
The models that were evaluated included a logistic regression classifier, a decision tree, a random forest and a light gradient boosting classifier. The results obtained with each model were the following:

| Model | Recall | Precision | F1-score | True positives | True negatives | False positives | False negatives | Average precision |
|---|---|---|---|---|---|---|---|---|
| Logistic regression | 0.807736 | 0.535371 | 0.643587 | 1235 | 2274 | 1067 | 294 | 0.635852 |
| Decision tree | 0.812999 | 0.528566 | 0.639675 | 1235 | 2273 | 1068 | 294 | 0.595305 |
| Random forest | 0.811653 | 0.534080 | 0.643995 | 1235 | 2273 | 1068 | 294 | 0.628662 |
| Light gradient boosting machine | 0.818842 | 0.534254 | 0.646322 | 1242 | 2251 | 1090 | 287 | 0.632229 |

As we can observe in the results, the performance of all models was quite similar. The model that was finally selected was the light gradient boosting machine classifier, since this model achieved the highest recall and F1-scores and the lowest number of false negatives.

Additional plots that were created to evaluate the performance of the light gradient boosting classifier are presented below:



# Conclusions

- The performance of all models was quite similar, although the light gradient boosting machine classifier achieved slightly better predictive performance. However, it needs to be considered that other features of the models such as its interpretability, easy implementation, and acceptability by the medical community may intervene in favour of a model that may not achieve the highest performance.

- The variables that were retained in the final model were age, gender, BMI, number of tumours, time from tumour diagnosis to surgery, tumour stage and tumour grade. Age, gender, BMI and number of tumours were not found to be associated with the probability of survival during the stepwise feature selection approach. However, they were retained in the model because age, gender and BMI are common confounders which are typically included in the models, and the number of tumours was useful to develop the visualisation. The variables that were found to have a higher impact on the probability of survival were time from tumour diagnosis to surgery, tumour stage and tumour grade, which emphasises the need for early diagnosis.

- Other variables that were not present in this dataset such as family history of colon cancer, diets high in calories and animal fat, alcohol consumption, obesity and carcinoembryonic antigen level, have all been found to significantly affect survival in colon cancer. The next update of the Simulacrum database was planned for this year, and it is expected that this new version will include more data and variables. Thus, it would be interesting to repeat this analysis including these other variables that affect survival, if they are available in the updated database.

- It would also be interesting to predict the probability of survival at 5 years and 10 years after diagnosis, which are frequently calculated in survival analyses, if these data are available in the updated Simulacrum database.

- With the increasing volume of clinical and real world data, machine learning tools may become increasingly important and they could potentially be used as tools to predict cancer survival, which could be translated into decision support tools in the medical domain to help clinicians make more meaningful decisions.