Utrecht University

Master Artificial Intelligence

&

LOGEX

Predicting cancer stage from healthcare claims.

Author:
Roan Oosenbrug

Daily supervisors:
Jan van der Eijk
Stefan Haan
First and Second examiner:
Krista Overvliet
David Terburg

July 6, 2020

Abstract

Purpose: Breast and colorectal cancer are among the most dominant types of cancer regarding incidence and mortality. Cancer staging is a critical part in the treatment of cancer patients, but is not represented in healthcare claims, while these claims are a rich source for finding more insight in cancer treatment. The purpose of this study is to predict cancer stage from healthcare claims, evaluating both model performance and predictor importance. Improvement on previous studies is attempted by broadening the range of predictors included by including indirectly linked activities and prescribed medicines, as well as classifying all 4 stages of cancer separately.
Methods: Data sets for the breast and colorectal cancer studies have been constructed by combining clinical patient data and care activity data from several different hospitals in the Netherlands. Multiple preprocessing steps have been applied to these data sets, including SMOTE and AENN to combat class imbalance. On these processed data sets, neural network, random forest, support vector machine and Super Learner models were trained to predict cancer stage from healthcare activities. These models were assessed based on AUC, sensitivity and specificity. Finally, predictor importance was determined via a combination of a model-agnostic interpretation method and a scoring system.
Results: The best performing model for breast cancer stage prediction was the random forest model with an AUC of 0.71. For the colorectal cancer study, the best performing model was the Super Learner model with feature selection, SMOTE and AENN, with an AUC of 0.61. These results show that the models have not been able to improve on results from previous studies. Predictor importance analysis showed a broad range of variables with high importance scores, including directly linked activities, indirectly linked activities as well prescribed medicines. These predictors however do not correspond to the treatment patterns described in the literature, as directly linked activities are underrepresented in the important predictors when compared to the literature.
Conclusion: This study has shown that using small and imbalanced data sets causes difficulties in constructing viable prediction models for predicting breast and colorectal cancer stage. However, including a broader range of predictors has been shown to be a possible improvement compared to previous studies. This motivates further research with larger, more balanced data sets.

List of abbreviations

Abbreviation	Meaning	Explanation
AI	Artificial intelligence	Branch of computer science which develops machines or applications capable of performing tasks that typically require human intelligence.
SEER-Medicare	Surveillance, Epidemiology and End-Results - Medicare	A large population-based data source combining clinical information with healthcare claims from America.
ICD-9-CM	International Classification of Diseases, Ninth revision, Clinical Modification	System of assigning codes to diagnoses and procedures.
HCPCS	Healthcare Common Procedure Coding System	Standardized code system necessary for medical providers to submit healthcare claims to Medicare.
CCS	Clinical Classification Software	Tool for clustering patient diagnoses and procedures into clinically meaningful categories.
TNM system	Tumor, lymph Nodes and Metastasis system	Globally recognised standard for classifying cancer stage.
GDPR	General Data Protection Regulation	EU law on data protection and privacy.
DICA	Ducth Institue for Clinical Auditing	Dutch non-profit organisation providing registrations for healthcare providers and insurers.
DBC	Diagnose Behandeling Combinatie	Code system used in the Netherlands assigned to healthcare activities.
LASSO	Least Absolute Shrinkage and Selection Operator	A regression analysis method that performs variable selection.
CART	Classification And Regression Tree	Classification model based on a binary tree structure with discrete and real numbered outcomes.
NN	Neural Network	Computational networks that are biologically inspired, consisting of nodes, connections and connection weights.

SVM	Support Vector Machine	Supervised learning model which use separating hyperplanes to distinguish between classes.
SMOTE	Synthetic Minority Oversampling Technique	Method of creating synthetic data points for minority class to re-balance data set.
ENN	Edited Nearest Neighbor	Method of determining whether a data point is noise and should be removed from a data set.
AENN	All-k Edited Nearest Neighbors	Method of filtering noise from data set.
AUC	Area Under the Curve	Measure of how well a prediction model can separate between two classes.
CT	Computer Tomogram	Method of using x-ray imaging for examination.
MRI	Magnetic Resonance Imaging	Method of imaging via magnetic fields for examination.
PET	Positron Emission Tomography	Method of imaging via positron emitting radioactive substances for examination.
SPECT	Single Photon Emission Computed Tomography	Method of imaging via gamma emitting radioactive substances for examination.

Table 1: *List of all abbreviations used in this paper*

Introduction

Cancer remains one of the most prominent diseases in the present time. In 2018, 119,923 new cancer cases were recorded in the Netherlands alone, as well as 42,286 individuals dying due to cancer. Out of all types of cancer, breast and colorectal cancer are among the most dominant regarding incidence and mortality. 16,209 people had been diagnosed with breast cancer, with 3,300 people dying to the disease. For colorectal cancer, a total number of 14,921 people had been diagnosed with the disease, with 6,442 people dying [1]. Due to the devastating nature of these diseases, the need for research in diagnosing and prediction is ever present.

A critical part in treating any type of cancer is the stage of the disease. Staging helps determining appropriate treatment, as well as unify terminology across medical practitioners. Furthermore, the stage of the disease has an enormous impact on survival rates. For example, 5-year survival rates for breast cancer patients change from 97,5% for stage 1 to 54,6% for stage 4 [2]. This shows that correctly identifying cancer stage is critical in any application.

Healthcare organisations are increasingly active in finding more insight in existing healthcare data. One of the biggest sources for healthcare data is healthcare insurance claims, which are referred to as healthcare claims. These claims are an extensive record of any healthcare activity from any patient. Examples are chemotherapy and breast conserving surgery, but also dispensation of claimable medicines such as capecitabine. This makes healthcare claims a rich source for potential machine learning or data mining techniques. Prior studies have suggested models using varying machine learning or data mining techniques to extract insights from healthcare claims. These models are however mostly either lacking in performance or in broader applicability. This leaves room for further research in accurate, broad models to gain new insight from healthcare claims.

The study presented in this paper will investigate the predictive power of healthcare claims for predicting cancer stage in breast and colorectal cancer. The aim of this study is to create prediction models viable for use in the medical domain. If these models prove to be of a high enough standard regarding predictive performance, they can be used to gain more insight in treatment patterns and costs, ultimately assisting in a more effective and efficient treatment for these cancer types.

To achieve the aim of this study, the following research question has been formulated:
What are the relevant predictors for predicting the stage of breast and colorectal cancer using healthcare claims?

To answer this research question, it is best to split this question into three sub-questions:

- *What are the relevant predictors for predicting the stage of breast and colorectal cancer?*
- *Which type of model predicts the stage of breast and colorectal cancer best?*

- *Is such a model of a high enough performance level to be viable for use in healthcare analysis?*

These sub-questions will be individually assessed, to eventually answer the main research questions.

Prediction of cancer incidence or stage from healthcare claims is not a new subject. A review of previous studies in this topic will give a helpful overview, informing the current study on best practices as well as indicating potential points of improvement.

Previous studies [3], [4], [5], [6], [7], [8], [9] have shown mixed results, with a wide variety of techniques used. When reviewing these studies, a number of components emerge that are consistent over all studies. All studies chose to focus on healthcare activities with ICD-9 codes for the corresponding cancer type as variables for their prediction models, mostly extending it with demographic characteristics. This shows the importance of including healthcare information directly linked with the cancer type in question. Furthermore, while performance varied, all models were able to achieve either competent levels of success across all performance metrics, or excellent performance on a subset of performance metrics with performance on the remaining metrics remaining sub-par. This makes the potential of a viable prediction model of cancer stage from healthcare claims apparent.

However, room for improvement is also evident. All studies except for Whyte et al. [6] chose not to look beyond healthcare activities with ICD-9 codes for the corresponding cancer types. It has already been proven that ICD-9 codes are an insufficient indicator of cancer stage [10]. Looking at a broader range of variables such as prescribed medicines and indirectly related healthcare activities might provide the prediction models more capabilities of distinguishing between cancer stages. This approach has already been applied in the study of Chubak et al. [11], where a prediction model was constructed for identifying second breast cancer events (recurrence and second breast primary tumors), with promising results. Furthermore, only Smith et al. [9] developed a prediction model for distinguishing between multiple cancer stages, choosing to look at stages 1/2, 3 and 4. All of the other studies chose to look at either incidence in general, or focusing solely on stage 4. This leaves room for a prediction model for all 4 stages of cancer. Finally, none of the studies achieved high performance across all metrics. While almost all models perform exceptionally well regarding specificity, these models drop off in performance when looking at either sensitivity or PPV. The study of Brooks et al. [7], [8] achieved a more balanced performance over all metrics, but with only competent levels of success. This indicates a room for improvement regarding results. The goal of this study is to improve on these points, by looking at all cancer stages and including healthcare activities both directly and indirectly linked to the cancer types, as well as prescribed medicine, which will hopefully improve sensitivity performance compared to the previous studies.

An overview of the studies taken into consideration for this study can be found in table 2.

Author	**Goal**	**Source**	**Model**	**Variables**	**Performance**
Nattinger et al.	Identify incidence breast cancer cases	SEER-Medicare database	Self designed rule system combined with logistic regression	Diagnosis or healthcare activities linked to breast cancer based on ICD-9-CM and HCPCS codes	Sensitivity: 80.26% Specificity: 99.9% PPV: 91.66% - 94.87%
Freeman et al.	Identify incidence of breast cancer cases	SEER-Medicare database	Logistic regression	Diagnosis or healthcare activities linked to breast cancer based on ICD-9-CM and HCPCS codes	Sensitivity: 90% Specificity: 99.86% PPV: 70%
Nordstrom et al.	Identify metastatic / stage 4 breast, lung, colorectal and prostate cancer cases	Oncology Services Comprehensive Electronics data warehouse combined with National Council for Prescription Drug Programs claims	CART models	Age, gender and healthcare activities, diagnoses and drugs indicated by either oncologists or ICD-9 codes	Sensitivity: 60%-81% Specificity: 75%-97% PPV: 75%-86%

Whyte et al.	Identify metastatic / stage 4 breast, lung and colorectal cancer cases	Impact Intelligence Oncology Management (IIOM) database and Optum Research Database	30 different generic and tumor specific algorithms	Age, gender and diagnosis or healthcare activities linked to breast, lung or colorectal cancer based on ICD-9-CM and HCPCS codes	Sensitivity: 53%-59% Specificity: 85%-99% PPV: 55%-82%
Brooks et al.	Identify stage 4 lung cancer cases	SEER-Medicare database	Super-learner algorithm incorporating logistic regression, random forests, generalized additive regression, classification trees and pruned classification trees	Demographic characteristics and diagnoses or healthcare activities linked to lung cancer based on ICD-9 codes	Sensitivity: 76%-78% Specificity: 77%-79%
Smith et al.	Predict cancer stage for breast cancer cases	SEER-Medicare database	One logistic regression model for predicting stage 4 vs 1-3, one logistic regression model for predicting stage 3 vs 1-2	Demographic characteristics and diagnoses or healthcare activities based on ICD-9 codes	Stage 4 vs 1-3 model: Sensitivity: 81% Specificity: 89% PPV: 24% Stage 3 vs 1-2 model: Sensitivity: 83% Specificity: 78% PPV: 98%

Table 2: *Studies in predicting incidence and stage of cancer in healthcare claims [3], [4], [5], [6], [7], [8], [9].*

Staging and characteristics of cancer types

To help validate the inclusion of any healthcare activity or prescribed medicine as a variable for predicting cancer stage, an overview will be given of the current staging process for both breast and colorectal cancer. Risk factors, symptoms and general trends for treating each cancer type will also be detailed, to get a general understanding of what factors are involved with each cancer type. This information is summarized in tables 7, 8 and 9.

Cancer staging

The stage of cancer is based on the TNM system, maintained by the American Joint Committee on Cancer (AJCC). The TNM system looks at three components: The size and extent of the tumor (T), whether the cancer is in the lymph nodes (N) and whether metastases (spread to other parts of the body) has occurred (M). A rundown of the TNM system is given in tables 3, 4, 5 and 6

Category	**Breast**	**Colorectal**
T0	No evidence of primary tumor	No evidence of primary tumor
T1	Tumor $\leq$ 2 cm	Tumor invades submucosa
T2	Tumor $>$ 2 cm and $\leq$ 5 cm	Tumor invades muscularis propria
T3	Tumor $>$ 5 cm	Tumor invades through the muscularis propria into pericolorectal tissues
T4	Tumor of any size with direct extension to the chest wall and/or to the skin (ulceration or skin nodules)	Tumor penetrates to the surface of the visceral peritoneum or directly invades or is adherent to other organs or structures

Table 3: *Categories of tumor size (T) in TNM classification [12], [13]*

Category	**Breast**	**Colorectal**
N0	No regional lymph node metastases	No regional lymph node metastases
N1	Metastases to movable ipsilateral level I, II axillary lymph node(s)	Metastases in 1–3 regional lymph nodes

N2	Metastases in ipsilateral level I, II axillary lymph nodes that are clinically fixed or matted; or in clinically detected ipsilateral internal mammary nodes in the absence of clinically evident axillary lymph node metastases	Metastases in 4 or more regional lymph nodes
N3	Metastases in ipsilateral infraclavicular (level III axillary) lymph node(s) with or without level I, II axillary lymph node involvement; or in clinically detected* ipsilateral internal mammary lymph node(s) with clinically evident level I, II axillary lymph node metastases; or metastases in ipsilateral supraclavicular lymph node(s) with or without axillary or internal mammary lymph node involvement S	-

Table 4: *Categories of lymph nodes (N) in TNM classification [12], [13]*

Category	**Breast**	**Colorectal**
M0	No distant metastases	No distant metastases
M1	Distant metastases	Distant metastases

Table 5: *Categories of metastases (M) in TNM classification [12], [13]*

Stage	**Breast**	**Colorectal**
1	- T1;N0;M0 - T0;N1(lymph node tumor size < 2 mm):M0 - T1;N1(lymph node tumor size < 2 mm);M0	- T1;N0;M0 - T2;N0;M0
2	- T0;N1;M0 - T1;N1;M0 - T2;N0;M0 - T2;N1;M0 - T3;N0;M0	- T3;N0;M0 - T4;N0;M0

3	- T3;N1;M0 - T4;Any N;M0 - Any T;N2;M0 - Any T;N3;M0	- Any T;N1;M0 - Any T;N2;M0
4	- Any T;Any N;M1	- Any T;Any N;M1

Table 6: *Cancer staging in TNM classification [12], [13]*

Breast cancer

The risk factors for breast cancer can be grouped into two groups: non-modifiable and environmental risk factors [14]. The first important non-modifiable factor is age, with breast cancer frequency being significantly higher with patients older than 45 years. Second, sex plays a role in developing breast cancer, with breast cancer only sporadically being diagnosed in men. A third non-modifiable factor is race, with Caucasian women for example having a higher frequency of breast cancer occurrence compared to Hispanics. Furthermore, familial susceptibility has been shown to be a risk factor for breast cancer, with for example the *BRCA1* and *BRCA2* genes being indicated as having their function disorder increase the occurrence of breast cancer. The final non-modifiable factor constitutes natural hormonal changes. A delay of menarche of 2 years for instance is associated with a higher risk of breast cancer. Environmental risk factors are described by lifestyle decisions. Consuming products with high levels of fat or chemical substances increase risk of breast cancer, and a low amount of physical activity having a similar increase risk. Finally, artificial hormonal changes also increase breast cancer risk, with oral hormonal menopause therapy being associated with higher frequency of breast cancer.

The symptoms of breast cancer can be grouped into three groups: Breast lump, non-lump breast symptoms and non-breast symptoms [15]. A breast lump is the most common symptom for breast cancer, making it a generally known symptom. Non-lump breast symptoms include nipple abnormalities (such as retraction and change in appearance), breast pain or skin abnormalities (such as rash, infection and swelling). Finally non-breast symptoms are more uncommon. These symptoms include an auxiliary lump, back pain, fatigue and weight loss.

Treatment of breast cancer differs per stage [16]. For stages 1 and 2, the most common treatment is breast conserving surgery followed by radiation therapy. If breast conserving surgery is not possible due to either medical considerations or personal preference, a mastectomy is performed. Most patients also undergo some form of adjuvant therapy, being either chemotherapy, endocrine therapy or tissue-targeted therapy. Patients with stage 3 breast cancer have to first undergo induction systemic therapies, in the form of either chemotherapy or endocrine therapy. The purpose of these induction therapies is to shrink the tumor to make surgery possible. If the tumor responds to the induction, a combination of breast conserving therapy and radiotherapy can be suggested. If the tumor

does not respond, a complete mastectomy can still be an option. For patients with stage 4 breast cancer the focus of the treatment shifts more to palliative treatment. Any type of surgery is uncommon at this stage. A combination of radiotherapy, chemotherapy and endocrine therapy is usually applied to relieve pain from bone complications.

Colorectal cancer

Several risk factors are associated with colorectal cancer, which again can be divided into two groups: non-modifiable and environmental risk factors [17]. Following the trend of breast cancer, age is again a dominant risk factor. Incidence rates for colorectal cancer are up to 50 times higher for patients between 60 and 79 years old compared to those younger than 40 years. Second, a personal history with adenomatous polyps or inflammatory bowel disease (IBD) increase the risk of colorectal cancer as well. Around 95% of colorectal cancer develop from such polyps, and the risk of developing colorectal cancer increases up to 20 fold for patients with IBD. Furthermore, a family history of colorectal cancer or adenomatous polyps is also linked to colorectal cancer incidence. Roughly 20% of patients who develop colorectal cancer have a family member with the same disease. Finally, inherited genes also play a role in developing colorectal cancer. For example, the *MLH1* and *MSH2* genes are responsible for developing hereditary nonpoly-posis colorectal cancer (HNPCC), which accounts for roughly 2-6% of all colorectal cancers. Similarly to the previously mentioned breast cancer risks, the environmental risk factors include smoking, dietary habits, a lack of exercise and alcohol consumption. Colorectal cancer on the other hand is predominantly environmentally caused. For example, a lack of exercise combined with being overweight is estimated to account for 25%-33% of colorectal cancer cases.

Colorectal cancer can present itself with multiple different symptoms [18]. The most frequent symptom of colorectal cancer is rectal bleeding. Within the rectal bleeding symptom, the nature of the bleeding can indicate the severity of the symptom. Dark blood usually requires a more urgent referral for further diagnosis. Next to rectal bleeding, a change of bowel movement is also a frequently occurring symptom. This includes diarrhoea and constipation. Furthermore, abdominal pain, weight loss and anaemia can also indicate colorectal cancer.

Treatment for colorectal again is differentiated for the different stages of the cancer [19]. For stage 1 and 2 patients, the most common treatment option is surgery. This can either be removing a part of the colon or rectum, called colectomy and proctectomy respectively, or removing specific polyps in the colon or rectum, called polypectomy. This potentially can be supplemented with adjuvant chemotherapy or radiotherapy. For stage 3 patients, polypectomy is usually no longer an option due to the size of the tumor. The most common treatment option is a colectomy or protectomy followed by adjuvant chemotherapy or radiotherapy. Patients with stage 4 colorectal cancer usually do not undergo surgery. The general treatment line for stage 4 patients is a com-

bination of chemotherapy, radiotherapy or targeted therapies. If the metastasis is of a small enough scale, surgery can still be an option. However, metastasis resection via surgery generally cannot be achieved.

Cancer type	**Risk factors**
Breast	Age, Sex, Race, Familial susceptibility, Natural hormonal changes, Diet, Low amount of physical activity, Artificial hormonal changes
Colorectal	Age, Personal history of adenomatous polyps or IBD, Family history, Inherited genes, Smoking, Diet, Lack of physical activity

Table 7: *Risk factors per cancer type*

Cancer type	**Symptoms**
Breast	Breast lump, Nipple abnormalities, Breast pain, Skin abnormalities, Auxiliary lump, Back pain, Fatigue, Weight loss
Colorectal	Rectal bleeding, Diarrhoea, Constipation, Abdominal pain, Weight loss, Anaemia

Table 8: *Symptoms per cancer type*

Cancer type	**Treatment Stage 1 and 2**	**Treatment Stage 3**	**Treatment Stage 4**
Breast	- Breast conserving therapy with radiation therapy - Mastectomy - Adjuvant chemotherapy, endocrine therapy or tissue-targeted therapy	- Endocrine therapy / chemotherapy followed by breast conserving surgery with radiation therapy - Endocrine therapy / chemotherapy followed by mastectomy	- Endocrine therapy - Chemotherapy - Radiation therapy
Colorectal	- Colectomy / proctectomy - Polypectomy - Adjuvant chemotherapty - Adjuvant radiation therapy	- Colectomy / protectomy - Adjuvant chemotherapy - Adjuvant radiation therapy	- Chemotherapy - Radiation therapy - Targeted therapy - Metastasis resection via surgery

Table 9: *Treatment per cancer type*

Reviewing the characteristics of each cancer type, it becomes apparent that the symptoms of each cancer type is varied and extensive. This indicates a broad source of reasons for a patient to undergo a healthcare activity when the patient has developed any of the cancer types. This supports the idea of not only including directly linked healthcare activities as variables for the prediction models, but also looking at indirectly linked healthcare activities.
Furthermore, treatment patterns do seem to indicate no significant differences between stage 1 and 2 for each of the cancer types. This would indicate that prediction models based on healthcare activities might struggle between differentiating between these stages. However, since healthcare claim data also includes additional information besides treatments such as medicine prescriptions, there is still a cautious potential for a prediction model to differentiate between all stages for both of the cancer types.

Methods

Data source

The data used for this study are processed by *LOGEX*, a healthcare analytics company in the Netherlands, and *MRDM*, a medical data processing company in the Netherlands. A total of 8 different hospitals located in the Netherlands deliver data to LOGEX and MRDM, and this data has been used in this study. The process of data delivery is visualized in figure 1. All data is handled and processed according to the GDPR. Both companies are working under a data processing agreement with the hospitals, all data is only processed for tasks which are specified in advance, all data is combined via a secure connection according to the NEN Norm 7512 and all data is always anonymised [20], [21].

The hospitals deliver two kinds of data: clinical patient data and care activity data. The clinical patient data is delivered to the *Dutch Institute for Clinical Auditing* (DICA) registries. This data contains medical history, characteristics and outcomes for patients. The data processing company then processes and validates this data delivery. The care activity data is delivered to the healthcare analytics company following the *DBC* structure [22]. This data contains the collection of healthcare claims for a hospital, which provides a list of healthcare activities. The healthcare analytics company then processes and validates this data delivery. The data processing company exports patient codes to the healthcare analytics company, so the care activity data can be linked to the patients. This care activity data linked to patients is then exported back to the data processing company, who then link that information back to the clinical patient data. This creates a data source of the patients medical information and the patients healthcare activities. This data source is finally stored on a cloud-based SQL server. Afterwards, the data processing company verifies this data source via a data-verification pipeline. Any possible error in the data, indicated by this verification pipeline, is then reported back to the corresponding hospital. This supports hospitals in maintaining data integrity, as well as helping them

get insight in their data.

This study will use the data source stored on the cloud-based server as the data for modelling. This predefined structure of medical information and healthcare activities per patient makes selecting patients based on cancer type and stage easy, as well as making the selection of healthcare activities easier, due to its linkage of cancer type and stage with healthcare activities by construction.

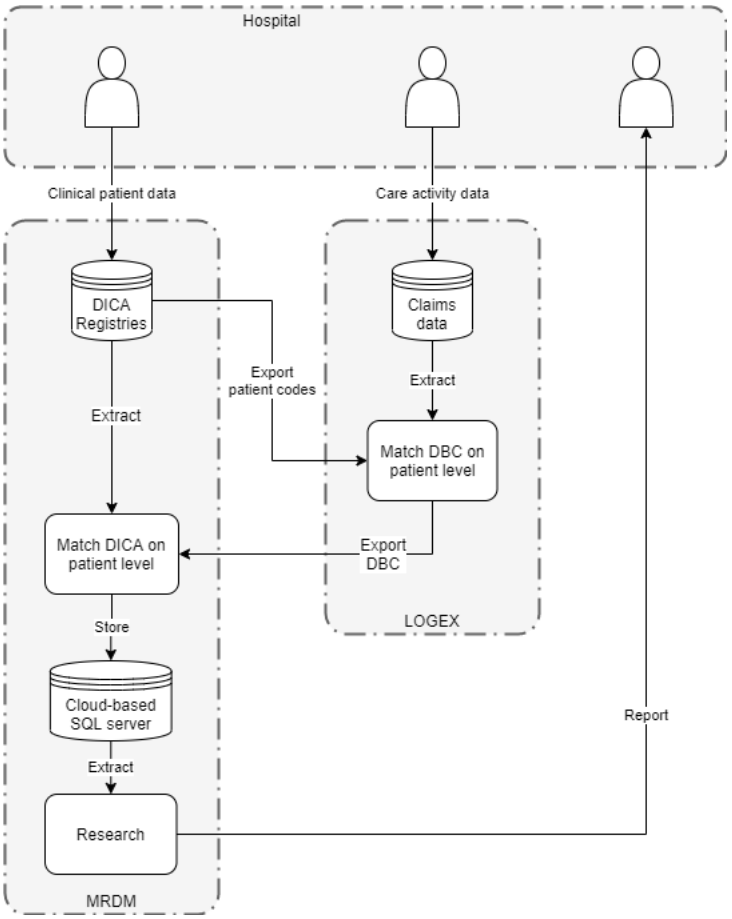

Figure 1: *Data processing workflow*

Model selection

Neural Network

Neural Networks (NN) are hierarchical networks designed to generate output from a combination of the input predictors [23]. They are inspired by biological neural networks, and are designed to mimic the structure and functionality of neurons. The network architecture consists of 3 parts: The input layer, one or

more hidden layers and an output layer. One layer consists of a set of nodes. The layers are connected via connections between individual nodes, and each connection has a corresponding weight. An illustration of this architecture is depicted in figure 2. All nodes have a computational model, which computes an output from a weighted sum from the input from the nodes of the previous layer of the network. This weighted sum needs to reach a minimum threshold, else the node will output 0.

The network produces an output by inputting the predictor values from one data point into the input layer, calculating the output for each layer via the computational models from each of the nodes until the output of the nodes of the output layer have been calculated. The output of this final layer is then used as the output of the model. For binary classification, the output layer will usually have either 1 node, with the value of that node deciding which label to output, or 2 nodes, one for each label.

Learning in a NN is done by error correction on the weights of connections. The learning procedure is done in 5 steps:

1. Initialize network with random weights for each connection
2. Input data point from training data, and calculate the network output
3. Calculate error based on the difference from the network output and the correct label of the data point
4. Update weights based on back propagation of error
5. Repeat step 2-4 until either all data points from training data have been inputted into the network or the network has converged

For a more thorough explanation of the structure and workings of NN's, the reader is encouraged to read the article from Jain et al. [23]. The NN model has been selected for this study due to its widespread use in research, its learning ability and adaptivity.

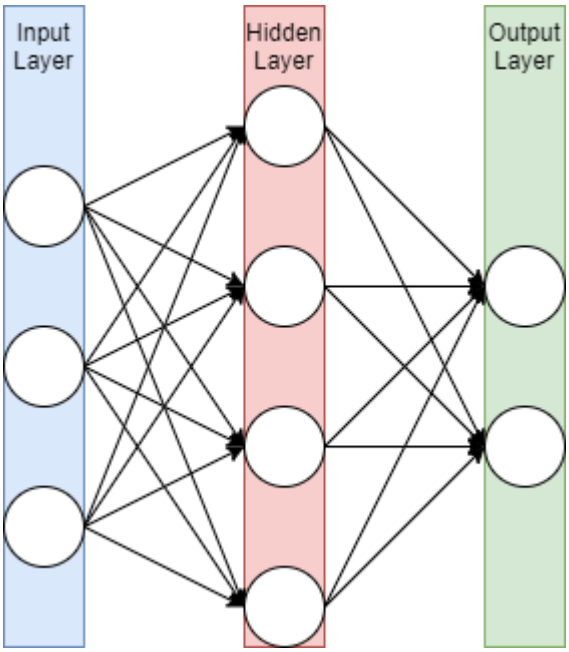

Figure 2: *An illustration of an example NN*

Random Forests

Random forests are a set of individual tree-like predictors, with the combination of this set resulting in a prediction model [24]. A single tree has a hierarchical architecture of nodes, with each node being a decision function on one of the predictors. Prediction of a single tree is done by traversing the tree from top to bottom, traversing each connection between nodes based on the decision function of the node. Prediction of a random forest model is done by letting each of the individual trees predict on the data point and taking the mode of all the predictions from the individual trees. An illustration of this architecture and prediction process is given in figure 3.

Construction of the individual trees is done via a process called *bagging*, where each tree samples with replacement from the training data. This results in a decrease of variance for the overall model, due to the decrease of correlation between individual trees. To further decrease variance of the overall model, each tree only gets a random subset of predictors to consider when determining the decision function for one node. This is done to decrease correlation between individual trees due to highly predictive predictors. The decision function is determined via a metric called the Gini index, which is a measure for a predictor of how often a random data point from the training data would be incorrectly classified if that predictor was used as a decision function.

For a more thorough explanation of Random Forests, the reader is referred to the paper of Breiman [24]. The Random Forests model is selected for this

study due to its widespread use in research, its power to handle high dimensional data sets and its low variability.

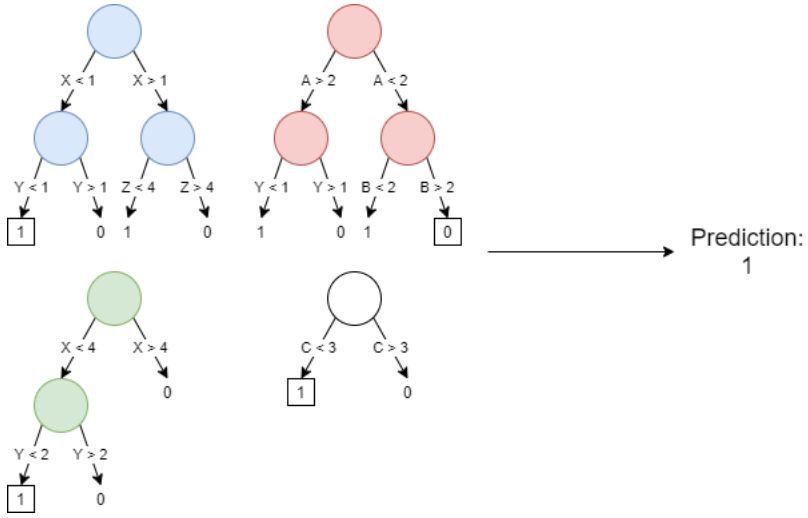

Figure 3: *An illustration of an example Random Forest*

Support Vector Machines

Support Vector Machines (SVM) are prediction models that use decision planes in a high dimensional feature space [25]. In the SVM models, data points are looked at as vectors in feature space. Most of the time these vectors are not linearly separable. To combat this problem, the SVM model transforms the vectors into a new space using a radial kernel function, making the vectors more easily separable. Once the vectors have been transformed into this new space, an optimal decision plane can be established. This is done by defining the plane as $w * x + b = 0$, where w is the normal vector of the hyper plane. x is the input vector and b is a scalar. Finding an optimum decision plane is done by maximizing the distance between the decision plane and the closest vectors of each class. These closest vectors are called *support vectors*. If the classes are not separable, the decision boundary is constructed by maximizing the distance to the support vectors, while minimizing the number of errors. An example of an SVM model is illustrated in Figure 4.

For a more thorough explanation of SVM models, the reader is referred to the paper of Cortes et al. [25]. The SVM model is selected for this study due to its performance in high dimensional data and its efficient procedure, making it ideal for real life application.

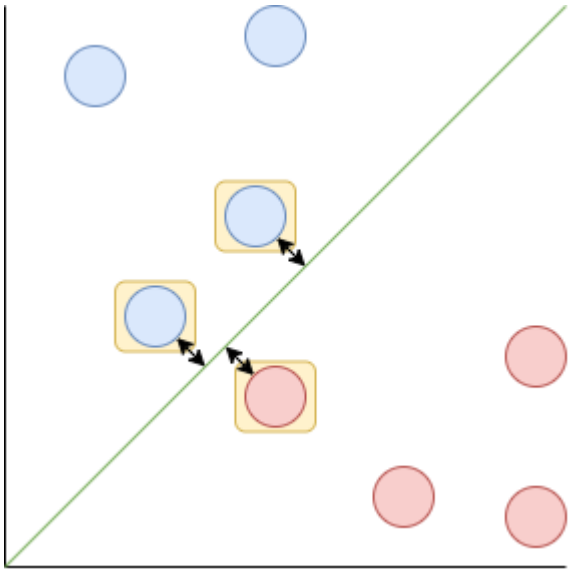

Figure 4: *An illustration of an example SVM Model. The green line indicates the decision plane, the circles indicated with yellow are the support vectors, and the black arrowed lines indicate the distance between the support vectors and the decision plane the model is trying to maximize*

Super Learner

A new trend in machine learning is combining multiple different models into one ensemble prediction model. This method is called super learning. One example of this method is the Super Learner model, proposed in the paper by Polley et al. [26]. The Super Learner model first individually fits a selection of provided models on the training data, and then creates a weighted combination on these models based on a 10-fold cross-validation on the training data. In this cross-validation, it establishes the weighted combination by minimizing the cross-validated risk.

Prediction by this model is done by first making predictions with each of the individual models, followed by using the weighted combination on the resulting predictions to derive a singular prediction. This method of construction and prediction has been proven to perform asymptotically well as the best possible weighted combination, as shown by Polley et al. [26].

The Super Learner model has been selected for this study due to its capabilities of improving on any singular prediction model, its proven capabilities in cancer stage prediction by Brooks et al. [7] and its ease of use.

Model construction

In order to construct the models, some model parameters need to be determined, as well as deciding which packages to use for implementation of the models. These choices will be detailed below. Any model parameter not mentioned below can be assumed to have the default value of the implementation. All models were implemented in R version 3.6.3 [27].

- **Neural Network**
  There is one important parameter to set for the NN models: the size of the hidden layer. A previous study by Wanas et al. has shown that a suitable amount of nodes for the hidden layer is ${}^2\log T$, where T is the number of training samples [28]. For the breast cancer study, there were a total of 1267 training samples, resulting in 6 nodes in the hidden layer for the NN models used in the breast cancer study. In the colorectal cancer study, there were a total of 1036 training samples. again resulting in 6 nodes in the hidden layer for the NN models. The NN models for this study were implemented with the *nnet* package [29].

- **Random Forests**
  Two important parameters in the Random Forests models are the number of trees constructed, and the number of predictors to consider when constructing a node. For the number of trees constructed, a previous study by Oshiro et al. has proven that a suitable amount of trees lies within a range of 64 and 128 [30]. Preliminary testing was done to determine what exact value should be chosen in this range. Random Forests models were created on the data with 64, 80, 96, 112 and 128 number of trees used, and performance of these models were compared. The best performing model was the model with 128 number of trees (Data not shown). Therefore, the value of number of trees was set to 128 for this study. For the number of predictors considered when constructing a node, a previous study by Svetnik et al. has shown that a suitable number of predictors to consider is $\sqrt{p}$, where p is the total amount of predictors [31]. For the breast cancer study, there were a total amount of 425 predictors, resulting in 21 random predictors considered in the models used for the breast cancer study. In the colorectal cancer study, there were a total of 522 predictors, resulting in 23 random predictors considered in the models used for the colorectal study. Finally, the Random Forests models constructed for this study were implemented with the *ranger* package [32].

- **Support Vector Machine**
  The SVM models in this study were implemented with the *glmnet* package [33].

- **Super Learner**
  The Super Learner models were implemented with the *SuperLearner* package [34].

Furthermore, since we are dealing with predicting all four stages of cancer, the prediction problem in this study is a multiclass classification problem. This needs to be considered when constructing the models, since the methods of dealing with multiclass classification are not consistent over all the models. To combat this issue, the *one-vs-all* scheme has been applied. This scheme consists of building separate binary models for each of the stages, setting all other stage class values in the train and test sets to a value of 0, and the stage class value of the stage in question to a value of 1. When predicting a new data point, all 4 separate binary models are run, and the stage is chosen as a prediction for which its corresponding binary model outputted the largest value. This scheme, while relatively simple, has been shown to be as accurate as any other multiclass classification scheme in a paper by Rifkin et al. [35]. This scheme has been chosen for this study for its simplicity, and the ability to separately analyse the prediction of any one of the stages when necessary.

Data preprocessing

Before model construction, multiple data preprocessing steps are applied to the raw data sets to facilitate model construction and optimize model trainability. A flowchart of these preprocessing steps is shown in Appendix figures A1 and A2. First, both the raw breast and colorectal cancer data sets are loaded from the cloud-based SQL server. These raw data sets consist of a large number of rows, where each row describes a single healthcare activity for a single patient. A single row contains a patient ID, the age of the patient, hospital code, healthcare activity code and T, N and M values for that patient. Important to note is that the these raw data sets already include only healthcare activities directly or indirectly related to the cancer types. This is done by filtering the activities on CCS codes when creating the data set, which are categorization codes for healthcare activities linking the activities to the overarching diagnoses [36]. The CCS codes used are 24 for breast cancer, and 14 and 15 for colorectal cancer. This helps selecting a broad range of activities which were used in either breast or colorectal cancer treatment, without selecting activities completely unrelated to these treatments. Secondly, all patients with invalid staging data are removed. Examples of invalid staging data are patients with NA values for T, N or M, or a combination of T, N and M values which do not correspond to the TNM staging system. Then, the T, N and M values are combined to a cancer stage value according to the TNM staging system, resulting in a value between 1 and 4. Next, data was transformed into a set of predictors and stages, where each row consists of the information for one patient. One row then contains the patient's age, the stage of the patient and a counter for all the healthcare activities seen over all patients. Finally, all the predictors are normalized to a range from 0 to 1. This is done to improve performance for neural network models [37].

After the construction of the predictors and stages set, a train and test set is constructed for both data sets by sampling 70% of the predictor and stages set as training data, and 30% as test data. To maintain a correct balance of stages

over both training and test data sets, the sampling was done proportional to the distribution of stages in the predictor and stages sets. After construction of the train and test sets, 4 separate train and test sets are created for the individual stages, since separate models will be created for the stages to facilitate the *one-vs-all* approach. These train and test sets are created by copying the original sets, and transforming all stage values to either a 1 if it is the corresponding stage, or a 0 if it is any of the other stages. Finally, all the tests sets are split into two separate sets, one containing the predictors and one containing the stages. This is done to facilitate model prediction on only the predictors.

Following the creation of the train and test data sets, feature selection is applied to the data. The LASSO feature selection method [38] is selected for this study, as it one of the most successful and widely used feature selection methods currently available. The LASSO method applies regularization, setting coefficients of non-important predictors to zero. For feature selection, the data sets are filtered to only include predictors which are non-zero after applying the LASSO method. The LASSO feature selection method was implemented using the *glmnet* package in R [33].

Finally, the distribution of stages is significantly imbalanced over both breast and colorectal data sets. These distributions will be detailed in the section *Patient Characteristics*. Since the models by definition are focused on reducing error rates, the imbalance of stages will have a negative impact on model performance. For example, in an extreme case a model could learn to ignore a stage if this stage is only present in 0.1% of data, since only predicting the opposite label would result in an error rate of only 0.1%. To combat this problem, two data preprocessing steps have been applied: Synthetic Minority Over-sampling Technique (SMOTE) and All-k Edited Nearest Neighbors (AENN).

Firstly, the SMOTE method [39] has been applied to re-balance the data sets. The SMOTE method first selects a random data point from the under-represented class. Then, it selects the 5 nearest neighbours of that data point with the same class. Next, it randomly chooses one of the 5 nearest neighbours. Finally, it creates a new synthetic data point by calculating the difference of the original data point and the selected neighbour, multiplying this difference by a random number between 0 and 1 and adding that result to the original data point. It repeats this process until enough synthetic data points have been created to result in an equally balanced data set. The SMOTE method has been implemented using the *smotefamily* package in R [40].

The negative consequence of SMOTE is the possible creation of data points which are not relevant for classification, or the creation of noise in the data. To combat this effect, the AENN method [41] has been applied. The AENN Method loops over all data points and uses the Wilson's Edited Nearest Neighbor Rule (ENN) [42] to remove data if necessary. The ENN rule selects the 5 nearest neighbours of a data point, and if the majority of the neighbours do not have the same class as the data point, the data point gets removed. The AENN method has been implemented using the *NoiseFiltersR* package in R [43]. This method in combination with SMOTE has been proven to be a good preprocessing procedure in a previous study by Batista et al. [44].

In this study, we will be evaluating the performance of the models in combination with one or more of these preprocessing steps. This results in 6 different models for each base model:

- Base Model
- Model + SMOTE
- Model + SMOTE + AENN
- Model + Feature Selection
- Model + Feature Selection + SMOTE
- Model + Feature Selection + SMOTE + AENN

Evaluation metrics

Model Performance

The evaluation metrics chosen for this study are sensitivity, specificity and area under the curve (AUC). These metrics have been chosen as they are almost exclusively used in the literature of predicting cancer stage, as well as being good indicators of model performance with imbalanced data, as opposed to e.g. accuracy. Since the models constructed use the *one-vs-all* scheme, the calculation and analysis of the evaluation metrics needs to be adapted accordingly. This study will analyse the evaluation metrics for each individual stage, as well as an average over all stages. For the calculation of the evaluation metrics for each stage, an approach similar to the *one-vs-all* scheme has been taken. Firstly, a prediction is made on the test set via the *one-vs-all* scheme, resulting in labels with values from 1 to 4. Secondly, a loop of 4 iterations is done over the prediction labels, one iteration for each stage. In each iteration, all the labels will be set to 0 if it does not correspond with the stage corresponding to this iteration, or set to 1 if it does correspond. Then, the evaluation metrics can be calculated with the edited prediction and the labels of the test data set. After all 4 iterations, the evaluation metrics for each stage have been calculated, and averages of these metrics over all stages can be calculated. This approach has been chosen instead of calculating the metrics on the predictions of the separate binary models themselves. If the metrics were calculated on the predictions of the separate binary models, then the metrics would no longer correspond to the performance of the overall model.

For answering the research sub-question *Is such a model of a high enough performance level to be viable for use in healthcare analysis?*, a threshold value needs to be established for judging the viability of the models. As discussed by Hosmer et al. [45], a general rule is that an AUC larger than 0.7 is acceptable, and an AUC larger than 0.8 is considered excellent. Since cancer staging is a very precise problem with little room for error, the threshold set in this study for a model to be viable is 0.8 or higher.

Finally, to enable valid comparison between models, each model has been constructed and evaluated 10 times, and the mean of each of the performance metrics has been calculated over these 10 runs. All results shown in the *Model performance* and *Relevant Predictors* subsections of the *Results* section are therefore mean values.

Predictor Importance

Since the prediction problem in this study is a multiclass classification problem, predictor importance needs to be established for each stage separately. In order to determine predictor importance for each stage over the different types of models, a combination has been made of a model-agnostic interpretation method and a scoring system.

The model-agnostic interpretation method used in this study is the *Model Reliance* method, proposed by Fisher et al. [46]. The package used to implement this method is *iml* [47]. This method was applied to all of the separate binary stage models to determine a list of all the predictors, ordered by importance. For each model, the *Model Reliance* procedure can be described as follows: Firstly, the labels in the test data are set to 0 and 1 for the corresponding stage as described in the previous paragraphs. Secondly, a base prediction is made on the test data, and the mean absolute error is calculated for this prediction. Thirdly, a predictor in the test data is permuted in a way so that the predictor is rendered uninformative. Then, a new prediction is made on the test data with the permuted predictor and the mean absolute error is again calculated. Next, the difference in mean absolute error between the prediction with the original test data and the prediction of the test data with the permuted predictor is calculated. This difference is then taken as a measure of predictor importance, with a higher difference meaning a more important predictor. After establishing the predictor importance of one predictor, the test data is reverted back to its original form. This process is then repeated until the predictor importance is determined for all the predictors. Once the predictor importance has been calculated for all the predictors, a list of the predictors ordered by predictor importance is made. This process is then repeated for all 4 stages, resulting in 4 ordered lists of predictors, one for each of the 4 stages for one model. Finally, this is then repeated for each of the models, which results in ordered lists of every predictor for each stage and type of model.

To combine these results into a predictor importance for the 4 stages over all models, a scoring method has been constructed. The method considers one stage at a time, and iterates over each of the ordered lists of predictors from that stage for every model. When looking at one ordered list, a score needs to be assigned to each of the predictors. The highest possible score is equal to the amount of predictors, and the lowest possible score is 1. The predictors are then traversed in order, assigning the highest score to the first predictor and then assigning a score to each predictor while descending in score by 1 at a time. This process is repeated for all of the models. Finally, the scores are summed up over all models, resulting in a definitive score for every predictor for one stage.

This is then repeated for all of the stages. Finally, as each models has been run 10 times, the means of the final scores are taken over all 10 runs. This scoring method was inspired by the ranking method used in the thesis of Fransen [48].

Using this combination of a model-agnostic interpretation method and a scoring method is beneficial for multiple reasons. It gives a good indication of predictors importance, it prevents one model having an excessive impact on the overall predictors importance and manages to provide a method of encapsulating predictors importance with one measure, while using different types of models.

Results

Patient Characteristics

A total of 1810 patients were included for the study on breast cancer, and a total of 1480 patients were included for the study on colorectal cancer. An overview of the patient characteristics for both the complete population and the population for each stage is given in tables 10 and 11.

	# of patients	Mean age	Median age
Total	1,810 (100%)	65.06	66
Stage 1	970 (53.6%)	65.77	67
Stage 2	706 (39.0%)	65.15	65
Stage 3	104 (5.7%)	58.10	55
Stage 4	30 (1.7%)	63.87	65

Table 10: *Patient characteristics in study population for breast cancer*

	# of patients	Mean age	Median age
Total	1,480 (100%)	72.03	74
Stage 1	455 (30.7%)	70.87	73
Stage 2	399 (27.0%)	74.24	76
Stage 3	538 (36.4%)	71.92	75
Stage 4	88 (5.9%)	69.65	71

Table 11: *Patient characteristics in study population for colorectal cancer*

Important to note is the imbalance in class distributions for both populations. For breast cancer, both stages 3 and 4 patients are heavily underrepresented (5.7% and 1.7% respectively), and stage 4 patients are heavily underrepresented in the colorectal patient population as well (5.9%). This poses a potential difficulty for modelling any machine learning model in these data sets

to properly predict all 4 stages, as well as it motivating the use of the SMOTE and AENN preprocessing techniques.

Model performance

Figure 5 shows the AUC averages over all stages for all the models. Further overviews of AUC results are shown in Appendix tables A3 and A4. The model which performed best for the breast cancer study , based on AUC, is the Random Forest model with an AUC value of 0.71. For the colorectal cancer study, the model which performed best is Super Learner with Feature Selection, SMOTE and AENN, with an AUC value of 0.61.

None of the models reached the AUC threshold of 0.8 to be considered viable. However, all models do perform better than chance (meaning an AUC value above 0.5), indicating that the models did indeed manage to learn some information about stage prediction. Interesting to see as well is the difference in performance between the two studies, with AUC averages for the breast cancer study ranging from 0.54 to 0.71, and AUC averages for the colorectal cancer study ranging from 0.52 to 0.61.

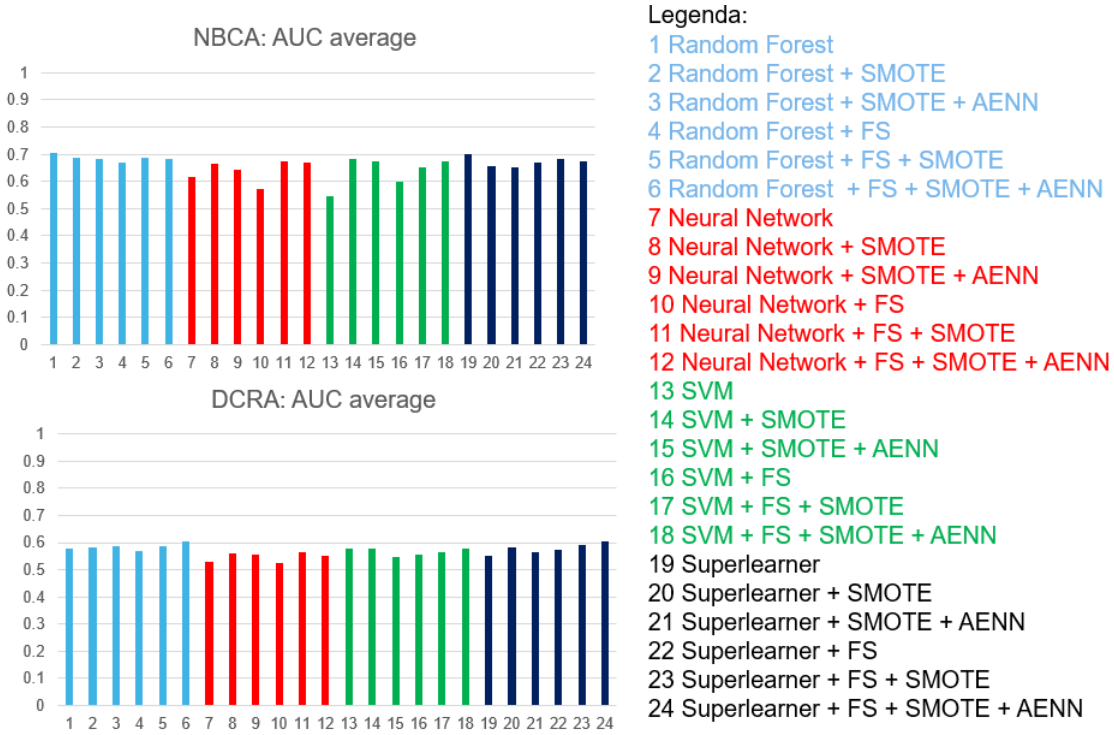

Figure 5: *The AUC averages of all the models for the breast cancer data (NBCA) and the colorectal cancer data (DCRA). For spacing purposes, feature selection has been abbreviated to FS.*

Figures 6 and 7 show the average sensitivity and specificity values over all stages for all models. Further overviews of sensitivity and specificity results are shown in Appendix tables A5, A6, A7 and A8. These figures show that performance in specificity was reasonable, and relatively consistent over all models. The average specificity value over all models was 0.84 for the breast cancer study, with values ranging from 0.78 to 0.90. For the colorectal cancer study, the average specificity value over all models was 0.78, with values ranging from 0.76 to 0.80. Performance in sensitivity however was considerably lower, as well as being more inconsistent over all models. The average sensitivity value over all models was 0.48 for the breast cancer study, with values ranging from 0.29 to 0.57. For the colorectal cancer study, the average sensitivity over all models was 0.36, with values ranging from 0.28 to 0.41. These poor sensitivity results help understand the inability of all models to reach the viability AUC threshold. Furthermore, this disparity between sensitivity and specificity performance is comparable to the disparity in performance of these metrics in previous studies described in the introduction. This shows that this study was unsuccessful in improving on this shortcoming of previous studies.

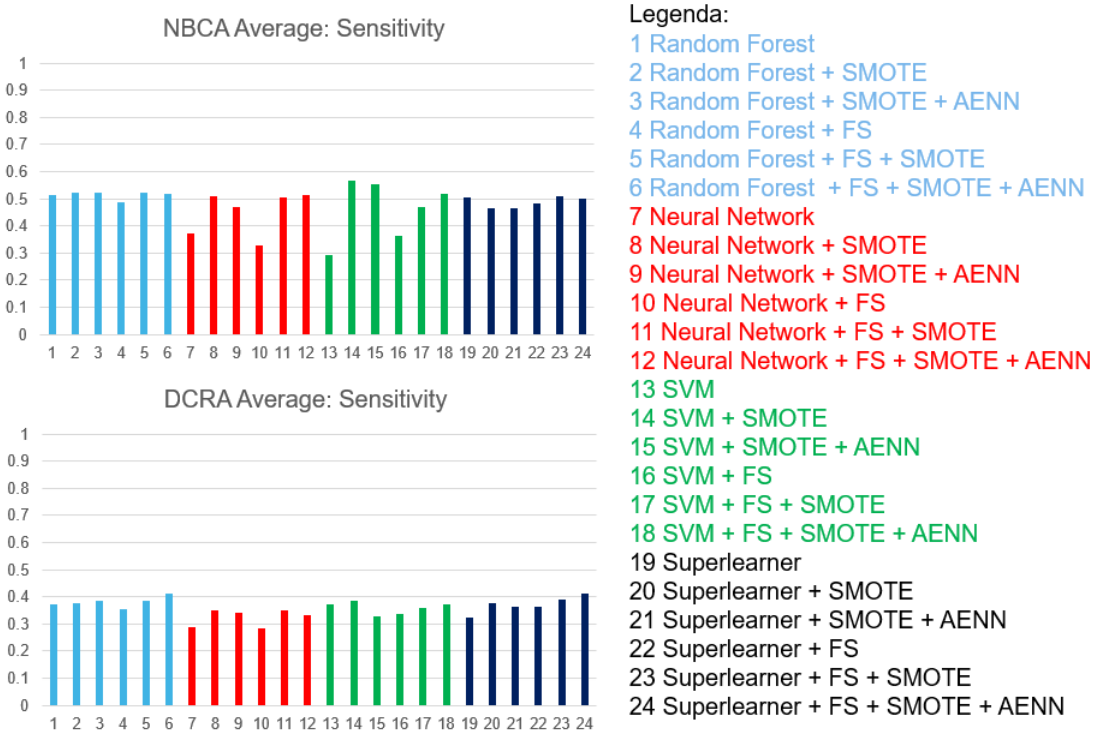

Figure 6: *The sensitivity averages of all the models for the breast cancer data (NBCA) and the colorectal cancer data (DCRA). For spacing purposes, feature selection has been abbreviated to FS.*

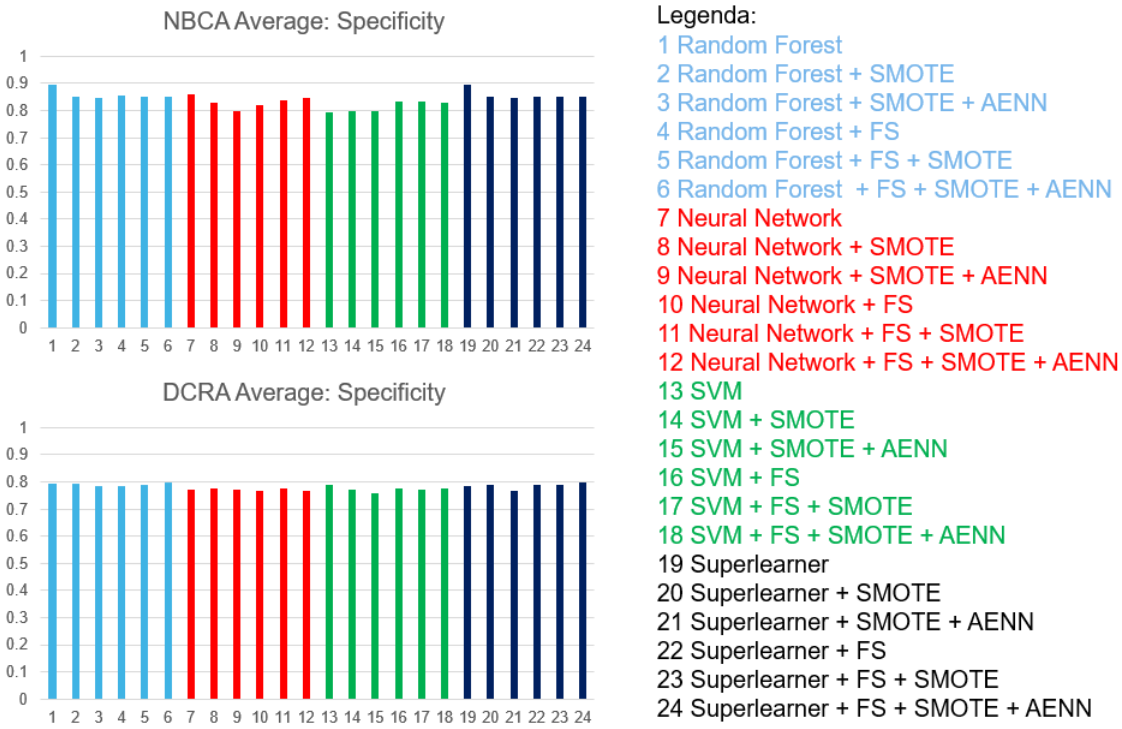

Figure 7: *The specificity averages of all the models for the breast cancer data (NBCA) and the colorectal cancer data (DCRA). For spacing purposes, feature selection has been abbreviated to FS.*

Relevant predictors

Appendix tables A9, A10, A11 and A12 detail the feature importance scores for the breast and colorectal studies, both for the models with and without feature selection. Descriptions for each of the predictors are detailed in appendix table A13. Only the top 10 predictors are displayed. The breast cancer study included 424 predictors, resulting in a highest possible score of 5088. Feature selection on the breast cancer data selected 50 predictors to be included in the models, resulting in a highest possible score of 600. For the colorectal cancer study, a total number of 521 predictors were included in the models, resulting in a highest possible score of 6252. Feature selection on the colorectal cancer data selected 61 predictors, resulting in a highest possible score of 732.
Figure 8 visualizes these results. 4 groups of healthcare activities have been identified: Diagnostic activities, invasive activities, medication prescriptions and a final grouping of any other activities. Figure 8 details how often an activity for any of these 4 groups occurred in the top 10 most important predictors (as detailed in appendix tables A9, A10, A11 and A12) for both the breast and colorectal cancer studies, with and without feature selection. This visualizes

the wide range of predictors identified as important in all of the models.

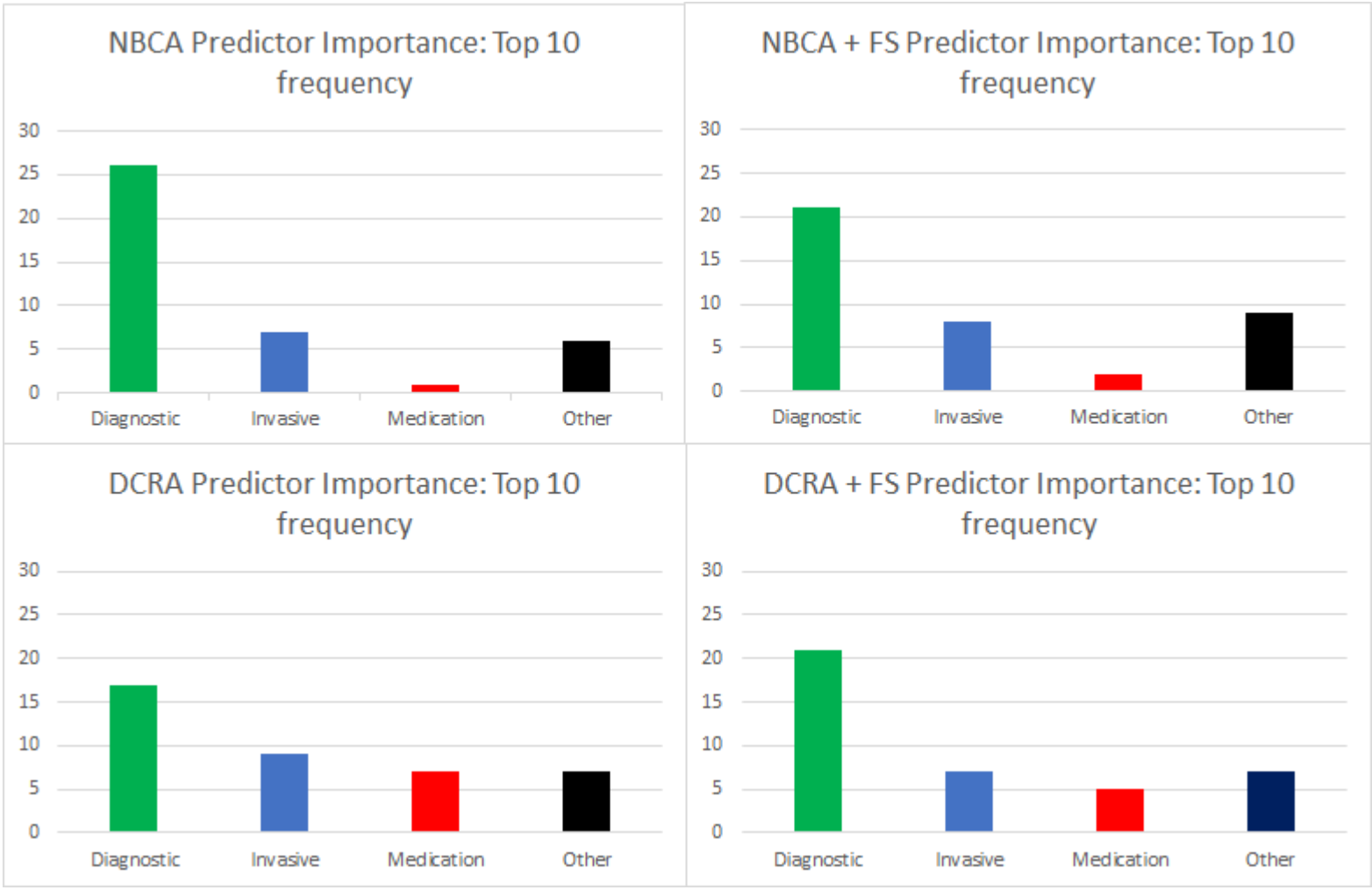

Figure 8: *The frequency of healthcare activities occuring the top 10 predictors for the breast cancer study (NBCA) and the colorectal cancer study (DCRA), both with and without Feature Selection (FS)*

Discussion

This study looked at what the relevant predictors for predicting the stage of breast and colorectal cancer are, which type of models performs best when predicting breast and colorectal cancer stages and whether these models would be viable to use in healthcare analysis. Each of the research sub-questions will now

be discussed individually.

What are the relevant predictors for predicting the stage of breast and colorectal cancer?

A wide range of predictors was found to be relevant for cancer stage prediction, including both directly linked healthcare activities as indirectly linked activities. These indirectly linked activities include medication prescriptions, diagnostic activities and secondary treatment for cancer patients.

These relevant predictors differ significantly from the treatment patterns described in the section *Staging and characteristics of cancer types*, specifically table 9. The models surprisingly focus more on diagnostic treatments, while the literature [16], [19] describes a significant difference in therapy itself. Furthermore, radiation therapy is missing as an important predictor across all models. This again is in contrast to the literature, which highlights the use of radiation therapy as a significant part of treatment for both cancer types. Additionally, while both models have at least some seemingly illogical predictors with a high score, the colorectal cancer study has a relatively large amount of illogical predictors with a high score (e.g. outpatient clinic activities, social work activities).

However, all models do show a logical pattern of predictor importance across the different stages. Reviewing the important predictors for each stage across models, one can see that predictors for stage 1 and 2 are generally concerning either general diagnostic activities, activities localizing the tumor itself or smaller operative activities. Important predictors for stage 3 and 4 generally describe activities that look for tumors over the whole body (meaning a search for metastasis), chemo / hormone therapy and larger operative activities. These patterns show that all models did establish logical predictors, indicating both the existence of viable information in the data sets, as well as the learning capabilities of these models. Furthermore, the important predictors across all models include both indirectly related activities (e.g. Treatment of wounds > 5 cm, Pathological investigation of simple biopsy or simple cytology, Ultrasound breast) and prescribed medicines (e.g. Herceptin, Paclitaxel, Avastin). This shows the benefit of not only including directly linked activities, but a broader range of variables which give more detailed information about the treatment of a patient.

Which type of model predicts the stage of breast and colorectal cancer best?

When looking at the best performing models, a random forest model has been shown to be the best performer when predicting breast cancer stage, with an AUC value of 0.71, and a Super Learner model with feature selection, SMOTE and AENN has been shown to be the best performer when predicting colorectal cancer stage, with an AUC value of 0.61. When looking at sensitivity and specificity, these two models as well as all other models performed significantly lower on sensitivity compared to specificity. This is in compliance with the results described in previous literature [3], [4], [5], [6], [7], [8], [9], where a

pattern was found across all previous studies of a relatively low performance on sensitivity. The average drop off in performance from sensitivity to specificity in the previous studies was 11.6%, while for this study it was 36.2% for the breast cancer study, and 42.1%. Important to note is the significantly worse decrease in this study compared to the previous studies. This can be explained by the difference in prediction problem. The prediction problem in this study was predicting all 4 stages of cancer separately for both breast and colorectal cancer, while the previous studies focused on a more general prediction problem, such as predicting incidence or predicting only 1 stage. This significantly increases the difficulty of prediction for the models, which results in a bigger decrease in performance from specificity to sensitivity.

Is such a model of a high enough performance level to be viable for use in healthcare analysis?

Since none of the models reached the AUC threshold value of 0.8, it can be concluded that none of the models are viable to use for health care analysis when trained on the data provided for this study. This inability for any of the models to reach the threshold performance value can be explained due to a multitude of limitations for this study.

Firstly, the data set size in this study was relatively small, which has a significant negative impact on model performance. While there is no definitive rule on the required data set size for achieving viable model performance, a comparison can be made to the data set sizes of the studies in previous studies. The mean data set size across the studies in previous studies is 15,882, ranging from 1,385 to 77,306. Comparing that to the data set size of this study (1,810 and 1,480 respectively), one can see that the data sets in this study are relatively small. Furthermore, the ratios between predictors and data points in the data sets are relatively large (424 predictors to 1,810 data points for the breast cancer study, 521 predictors to 1480 data points for the colorectal cancer study). Again, no definitive rule exists for a sufficient ratio between predictors and data points to achieve viable model performance. However, one estimate was made in an article by Haldar [49], where a general rule of thumb was established. This rule of thumb states that the ratio between predictors to data points in the training data should be 1:10. For the breast cancer study, this would result in a minimum of 4240 data points needed, and for the colorectal cancer study it would result in a minimum of 5210 data points needed. This is significantly more than the data set size available for this study, which partly explains the models not achieving viable performance.

Secondly, the data sets used in this study were imbalanced, with the most extreme cases being an imbalance level of 3:181 for stage 4 in the breast cancer and an imbalance level of 11:185 for stage 4 in the colorectal cancer studies. The study of Somasundaram et al.[50] has shown that data imbalance has a negative impact on model performance, as well as showing that even after constructing an effective algorithm with specific steps for dealing with imbalanced data, model performance is still hampered by the imbalance. This partly explains the mod-

els not achieving viable performance, even after applying SMOTE and AENN for dealing with imbalance in the data sets.

Moreover, the data used in this study consists of healthcare claims. These claims are generated purely for billing purposes, not for scientific purposes. The data therefore is susceptible to incomplete, unverified or erroneous data points [51]. While preprocessing steps have been included to remove invalid data points, the lacking scientific quality of the data also partly explains the lacking model performance.

Furthermore, while the predictor importance analysis shows the impact of including medication prescription in modelling cancer stage prediction, not all medication prescriptions are included in the data. The healthcare activity data provided for this study does not include non-expensive prescription medication, as these prescriptions do not need to be billed, and therefore do not appear as a healthcare claim. This indicates a proportion of missing data when incorporating medicine prescriptions in the prediction models.

Finally, some healthcare activities are billed 'in-house', meaning that the billing of these activities are followed within the hospital, and not via a healthcare provider. This results in these activities not being present in the healthcare activity data provided for this study. The most glaring example of this is radiotherapy activity. While the literature describes radiotherapy as one of the most common activities in treatment of both breast and colorectal cancer, the healthcare activities related to these treatments are not identified as one of the most important predictors for any of the models. Some further analysis found that these activities were not present in the data at all, due to the in-house billing of these activities. This shows that the data is missing some healthcare activities, which results in a part of the treatment of a patient potentially being absent in the data sets.

This study has two further limitations, which do not impact model performance, but are still important to highlight. Firstly, all the models have been trained on data available at the time of this study. If in the future new treatments will be found and applied for cancer treatment, or new treatment guidelines will be applied for cancer treatment, the models will have to be retrained and re-evaluated to incorporate these changes. Secondly, the models have been trained on data from a select set of regional Dutch hospitals, without teaching or academic foundations. Treatment patterns and guidelines might differ between hospitals in the Netherlands, or even internationally. If these models would have to be applied in a broader case, further research must be done to evaluate model performance on data sets from different hospitals.

Despite the prediction models not achieving viable performance levels, the methodology in this study has been shown to be a possible improvement on previous studies and is a reasonable subject for further research. With the broader range of predictors considered in the models being reflected in the most important predictors, combined with the proven potential of similar methods in previous studies and the described limitations of the data set, a reasonable suggestion can be made that the lacking performance of the predictions is likely due to the inadequate data sets provided, not due to the methodology. Further

research is therefore needed to support this suggestion. A follow-up study with larger, more balanced data sets could provide a significant increase in performance. Furthermore, a follow-up study applying this methodology to data sets from different hospitals, possibly from different countries, is needed to show the applicability of this study to different populations of patients. Moreover, while a number of arguments have been given to support the selection of neural network, random forest, support vector machine and Super Learner models, a different selection of models can possibly provide different insights and improve prediction performance. Finally, while predicting each stage separately provides the most detailed classification of cancer stage, constructing models on classes of cancer stage instead. For example, predicting stage 1 and 2 vs stage 3 vs stage 4 might provide better prediction performance. After further analysis on the results of the breast cancer study, an average of 97.1% of the erroneous predictions for stage 1 across all models was caused by predicting stage 2, and an average of 90.8% of the erroneous predictions for stage 2 across all models was caused by predicting stage 1. This indicates that grouping these two stages into one class could provide better performing models, at the cost of clinical insight.

Broadening the scope of this study, further research on similar cases is also an interesting field for research. The generalizability of this methodology to predict cancer stage can be investigated by applying the methodology to different types of cancer. Furthermore, further research can look into not only predicting cancer stage, but also classifying specific types of cancer. (e.g. HER2 positive cancer).

In conclusion, this study has shown that using small and imbalanced data sets causes difficulties in constructing viable prediction models for predicting breast and colorectal cancer stages. However, including a broader range of predictors has been shown to be a possible improvement compared to previous studies. This motivates further research with larger, more balanced data sets. Cancer stage prediction remains a difficult but interesting topic of research, and studies like these will help in the development of viable prediction models that can provide insight in treatment patterns and costs, as well as assisting in establishing a more effective and efficient treatment for cancer patients.

References

[1] GLOBOCAN, *Cancer statistics of the Netherlands 2018*, 2018. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/populations/528-the-netherlands-fact-sheets.pdf.

[2] M. Sant, C. Allemani, R. Capocaccia, T. Hakulinen, T. Aareleid, J. W. Coebergh, M. P. Coleman, P. Grosclaude, C. Martinez, J. Bell, J. Youngson, F. Berrino, A. Kupp, G. Hedelin, G. Chaplain, C. Exbrayat, B. Tretarre, J. Mace-Lesech, A. Danzon, M. Mercier, N. Raverdy, E. Artioli, M. Federico, A. Barchielli, E. Paci, G. Gatta, P. Crosignani, D. Speciale, M. R. Ruzza, E. Frassoldi, A. Verdecchia, L. Gafa, R. Tumino, M. La Rosa, A. Voogd, and E. M. Williams, "Stage at diagnosis is a key explanation of differences in breast cancer survival across Europe," *International Journal of Cancer*, vol. 106, no. 3, pp. 416–422, Sep. 2003. DOI: 10.1002/ijc.11226.

[3] A. B. Nattinger, P. W. Laud, R. Bajorunaite, R. A. Sparapani, and J. L. Freeman, "An Algorithm for the Use of Medicare Claims Data to Identify Women with Incident Breast Cancer," *Health Services Research*, vol. 39, no. 6p1, pp. 1733–1750, 2004. DOI: https://doi.org/10.1111/j.1475-6773.2004.00315.x.

[4] B. L. Nordstrom, J. L. Whyte, M. Stolar, C. Mercaldi, and J. D. Kallich, "Identification of metastatic cancer in claims data," *Pharmacoepidemiology and Drug Safety*, vol. 21, no. SUPPL.2, pp. 21–28, May 2012. DOI: 10.1002/pds.3247.

[5] J. L. Freeman, D. Zhang, D. H. Freeman, and J. S. Goodwin, "An approach to identifying incident breast cancer cases using Medicare claims data," Tech. Rep., 2000, pp. 605–614. DOI: https://doi.org/10.1016/S0895-4356(99)00173-0.

[6] J. L. Whyte, N. M. Engel-Nitz, A. Teitelbaum, G. Gomez Rey, and J. D. Kallich, "An Evaluation of Algorithms for Identifying Metastatic Breast, Lung, or Colorectal Cancer in Administrative Claims Data," *Medical Care*, vol. 53, no. 7, pp. 49–57, 2015. DOI: https://doi.org/10.1097/MLR.0b013e318289c3fb.

[7] G. A. Brooks, S. L. Bergquist, M. B. Landrum, S. Rose, and N. L. Keating, "Classifying Stage IV Lung Cancer From Health Care Claims: A Comparison of Multiple Analytic Approaches," JCO Clinical Cancer Informatics, Tech. Rep., 2019. DOI: 10.1200/CCI.18.00156.

[8] S. L. Bergquist, G. A. Brooks, N. L. Keating, M. B. Landrum, and S. Rose, "Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data," Tech. Rep., 2017.

[9] G. L. Smith, Y. T. Shih, S. H. Giordano, B. D. Smith, and T. A. Buchholz, "A method to predict breast cancer stage using Medicare claims," *Epidemiol Perspect Innov*, vol. 7, no. 1, 2010. DOI: 10.1186/1742-5573-7-1.

[10] G. S. Cooper, Z. Yuan, K. C. Stange, S. B. Amini, L. K. Dennis, and A. A. Rimm, "The Utility of Medicare Claims Data for Measuring Cancer Stage," *Medical Care*, vol. 37, no. 7, pp. 706–711, 1999. DOI: https://doi.org/10.1097/00005650-199907000-00010.

[11] J. Chubak, O. Yu, G. Pocobelli, L. Lamerato, J. Webster, M. N. Prout, M. Ulcickas Yood, W. E. Barlow, and D. S. Buist, "Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer," *Journal of the National Cancer Institute*, vol. 104, no. 12, pp. 931–940, Jun. 2012. DOI: 10.1093/jnci/djs233.

[12] American Cancer Society, *Breast Cancer Staging Poster*, 2009. [Online]. Available: https://cancerstaging.org/references-tools/quickreferences/Documents/BreastMedium.pdf.

[13] ——, *Colon and Rectum Cancer Staging Poster*, 2009. [Online]. Available: https://cancerstaging.org/references-tools/quickreferences/documents/colonmedium.pdf.

[14] M. Kamińska, T. Ciszewski, K. Łopacka-Szatan, P. Miotła, and E. Starosławska, "Breast cancer risk factors," *Przeglad Menopauzalny*, vol. 14, no. 3, pp. 196–202, 2015. DOI: 10.5114/pm.2015.54346.

[15] M. M. Koo, C. von Wagner, G. A. Abel, S. McPhail, G. P. Rubin, and G. Lyratzopoulos, "Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis," *Cancer Epidemiology*, vol. 48, pp. 140–146, Jun. 2017. DOI: 10.1016/j.canep.2017.04.010.

[16] K. L. Maughan, M. A. Lutterbie, and P. S. Ham, "Treatment of Breast Cancer," Tech. Rep. 11, 2010, pp. 1339–1346. [Online]. Available: www.aafp.org/afpAmericanFamilyPhysician1339.

[17] F. A. Haggar and R. P. Boushey, "Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors," *Clinics in Colon and Rectal Surgery*, vol. 22, no. 4, pp. 191–197, 2009. DOI: 10.1055/s-0029-1242458.

[18] W. Hamilton and D. Sharp, "Diagnosis of colorectal cancer in primary care: The evidence base for guidelines," *Family Practice*, vol. 21, no. 1, pp. 99–106, Feb. 2004. DOI: 10.1093/fampra/cmh121.

[19] E. J. Kuipers, W. M. Grady, D. Lieberman, T. Seufferlein, J. J. Sung, P. G. Boelens, C. J. Van De Velde, and T. Watanabe, "Colorectal cancer," *Nature Reviews Disease Primers*, vol. 1, Nov. 2015. DOI: 10.1038/nrdp.2015.65.

[20] LOGEX, *Privacystatement Logex*, 2019. [Online]. Available: https://www.logex.com/nl/privacy.

[21] MRDM, *Privacystatement MRDM*. [Online]. Available: https://mrdm.nl/privacystatement-mrdm/.

[22] Nederlands Zorgautoriteit, *Handleiding dbc-systematiek*, 2018.

[23] A. Jain, J. Mao, and K. Mohiuddin, "Artificial neural networks: a tutorial," *Computer*, vol. 29, no. 3, 1996. DOI: 10.1109/2.485891.

[24] L. Breiman, "Random Forests," Tech. Rep., 2001, pp. 5–32. DOI: https://doi.org/10.1023/A:1010933404324.

[25] C. Cortes, V. Vapnik, and L. Saitta, "Support-Vector Networks," Tech. Rep., 1995, pp. 273–297. DOI: 10.1007/BF00994018.

[26] E. C. Polley and M. J. Van Der Laan, "Super Learner In Prediction," U.C. Berkeley, Tech. Rep., 2010. [Online]. Available: http://biostats.bepress.com/ucbbiostat/paper266.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, 2017. [Online]. Available: https://www.R-project.org/.

[28] N. Wanas, G. Auda, M. Kamel, and F. Karray, "On the optimal number of hidden nodes in a neural network," in *Conference Proceedings. IEEE Canadian Conference on Electrical and Computer Engineering*, Waterloo, 1998. DOI: 10.1109/CCECE.1998.685648.

[29] B. Ripley and W. Venables, *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*, 2016. [Online]. Available: https://CRAN.R-project.org/package=nnet.

[30] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7376 LNAI, 2012, pp. 154–168, ISBN: 9783642315367. DOI: 10.1007/978-3-642-31537-4{_}13.

[31] V. Svetnik, A. Liaw, and C. Tong, "Variable Selection in Random Forest with Application to Quantitative Structure-Activity Relationship," in *Proceedings of the 7th Course on Ensemble Methods for Learning Machines*, 2004. [Online]. Available: https://www.researchgate.net/publication/228572061.

[32] M. N. Wright, S. Wager, and P. Probst, *ranger: A Fast Implementation of Random Forests*, 2019. [Online]. Available: https://CRAN.R-project.org/package=ranger.

[33] J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, N. Simon, and J. Qian, *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, 2019. [Online]. Available: https://cran.r-project.org/web/packages/glmnet/index.html.

[34] E. Polley, E. LeDell, C. Kennedy, S. Lendle, and M. van der Laan, *SuperLearner: Super Learner Prediction*, 2019. [Online]. Available: https://CRAN.R-project.org/package=SuperLearner.

[35] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification," Tech. Rep., 2004, pp. 101–141.

[36] Healthcare Cost and Utilization Project (HCUP), *Clinical Classifications Software (CCS) for ICD-10-PCS (beta version)*. [Online]. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp.

[37] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Transactions on Nuclear Science*, vol. 44, no. 3 PART 3, pp. 1464–1468, 1997, ISSN: 00189499. DOI: 10.1109/23.589532.

[38] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010. DOI: https://doi.org/10.18637/jss.v033.i01.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. DOI: https://doi.org/10.1613/jair.953.

[40] S. Wacharasak, *smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE*, 2019. [Online]. Available: https://CRAN.R-project.org/package=smotefamily.

[41] I. Tomek, "An Experiment with the Edited Nearest-Neighbor Rule," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-6, no. 6, pp. 448–452, 1976. DOI: https://doi.org/10.1109/tsmc.1976.4309523.

[42] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, pp. 408–421, 1972. DOI: 10.1109/TSMC.1972.4309137.

[43] P. Morales, J. Luengo, L. Garcia, A. Lorena, A. de Carvalho, and F. Herrera, *NoiseFiltersR: Label Noise Filters for Data Preprocessing in Classification*, 2016. DOI: https://CRAN.R-project.org/package=NoiseFiltersR.

[44] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004. DOI: https://doi.org/10.1145/1007730.1007735.

[45] D. Hosmer, S. Lemeshow, and R. Sturdivant, *Applied Logistic Regression*, 3rd. Chicester: Wiley, 2013, p. 528, ISBN: 0470582472.

[46] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-760.html.

[47] C. Molnar and P. Schratz, *iml: Interpretable Machine Learning*, 2020. [Online]. Available: https://cran.r-project.org/web/packages/iml/index.html.

[48] L. X. Fransen, "Predicting length of stay, discharge destination and mortality of patients with hip fractures," PhD thesis, Utrecht University, 2019.

[49] M. Haldar, *How much training data do you need?* 2015. [Online]. Available: https://medium.com/@malay.haldar/how-much-training-data-do-you-need-da8ec091e956.

[50] A. Somasundaram and U. Srinivasulu Reddy, "Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data," in *ICRECT 2016*, 2016.

[51] L. I. Iezzoni, "Assessing Quality Using Administrative Data," Tech. Rep. 2, 1997, pp. 666–674. DOI: 10.7326/0003-4819-127-8{_}Part{_}2-199710151-00048.

[52] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN," Stanford University, Stanford, Tech. Rep., 1975, pp. 303–320. DOI: https://doi.org/10.1016/0010-4809(75)90009-9.

[53] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism: Clinical and Experimental*, vol. 69, S36–S40, Apr. 2017. DOI: 10.1016/j.metabol.2017.01.011.

[54] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, pp. 1–16, 2017. DOI: 10.4236/jilsa.2017.91001.

[55] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015. DOI: 10.1016/j.csbj.2014.11.005.

Appendix

A0: Relevance to AI

Artificial Intelligence techniques have been developed and used in the medical domain for decades. One of the first famous AI systems in the medical domain was the expert system *MYCIN*, developed in the 1970's [52]. This has since developed in a multitude of branches. Examples are novel therapeutic target discoveries using enhanced Markov clustering, complex ecosystems for treating chronic mental diseases using multi-agent systems and carebots being used in the delivery of care [53].
Diagnosing and predicting diseases specifically has been a topic of interest for many studies [54], [55]. These studies have shown the potential of using AI techniques for prediction and diagnoses of a multitude of diseases. This study could potentially add to this list of successful AI applications in the medical domain, both in practical use as well as knowledge discovery.

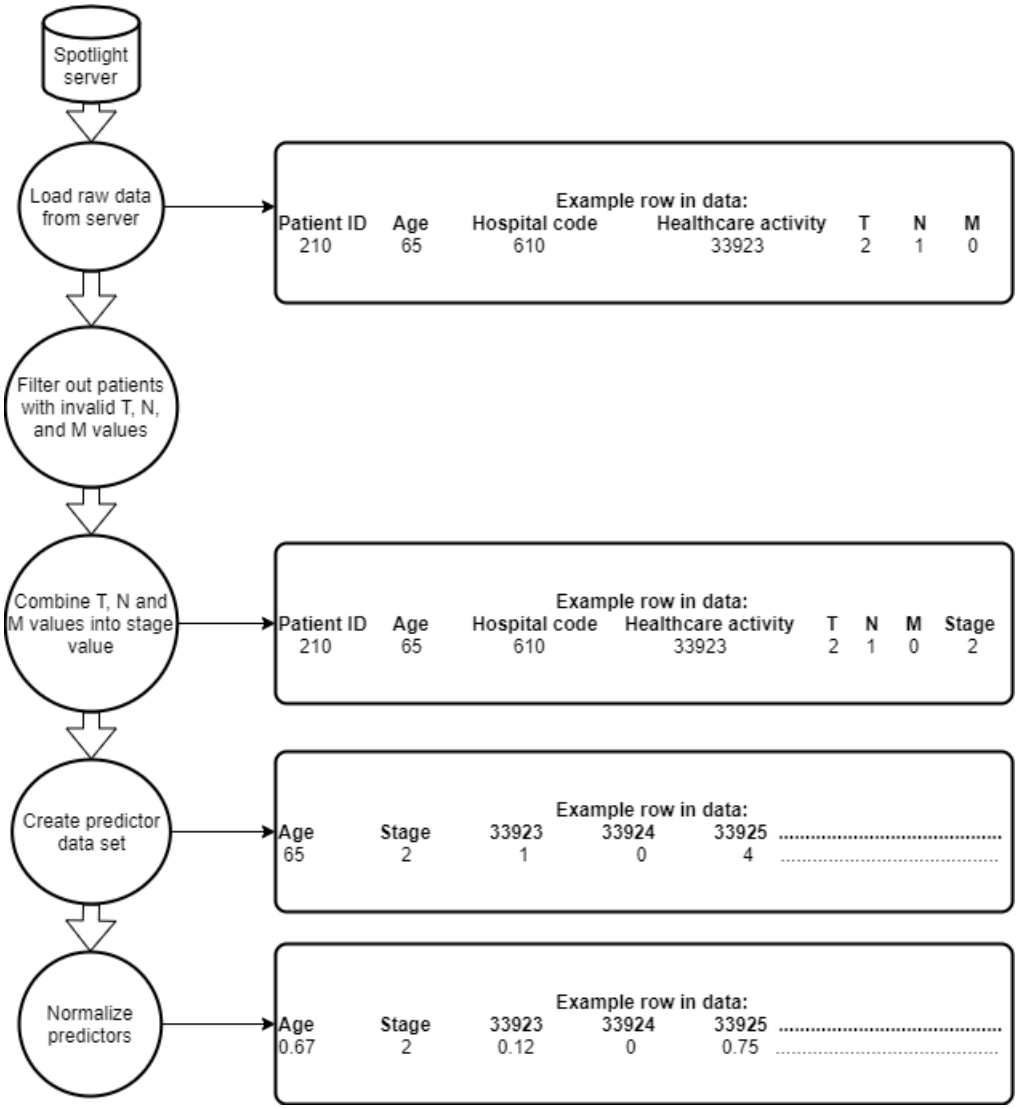

Figure A1: *Part 1 of a flowchart of the preprocessing steps used in this study*

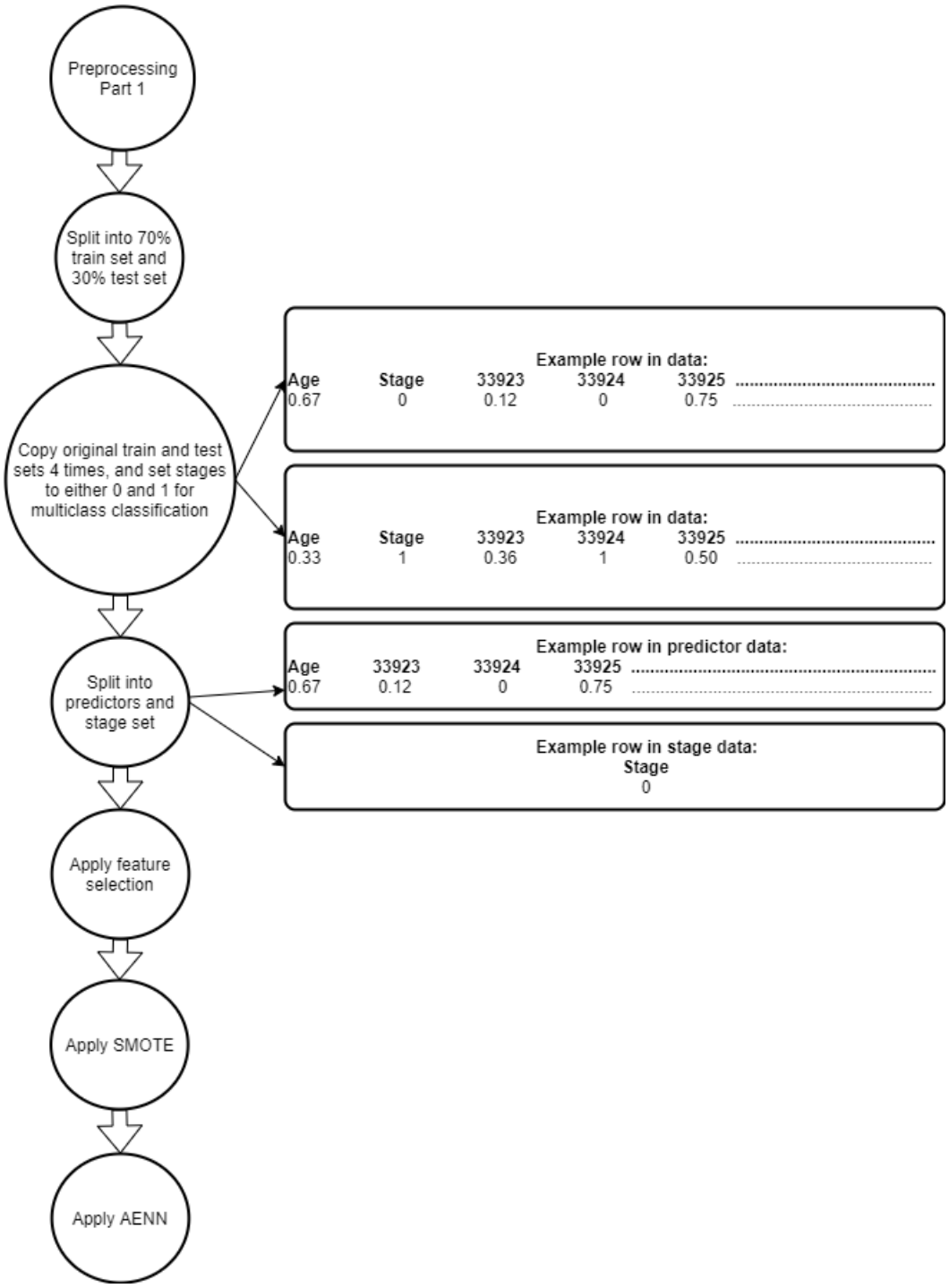

Figure A2: *Part 2 of a flowchart of the preprocessing steps used in this study*

Model	Stage 1	Stage 2	Stage 3	Stage 4	Average
Random Forest	0.83	0.77	0.53	0.69	0.71
Random Forest + SMOTE	0.73	0.68	0.58	0.75	0.69
Random Forest + SMOTE + AENN	0.72	0.66	0.61	0.75	0.68
Random Forest + Feature Selection	0.74	0.68	0.57	0.70	0.67
Random Forest + Feature Selection + SMOTE	0.74	0.68	0.59	0.75	0.69
Random Forest + Feature Selection + SMOTE + AENN	0.73	0.65	0.59	0.75	0.68
Neural Network	0.75	0.72	0.50	0.50	0.62
Neural Network + SMOTE	0.69	0.61	0.61	0.75	0.67
Neural Network + SMOTE + AENN	0.67	0.57	0.68	0.65	0.64
Neural Network + Feature Selection	0.66	0.64	0.50	0.50	0.57
Neural Network + Feature Selection + SMOTE	0.71	0.63	0.68	0.67	0.67
Neural Network + Feature Selection + SMOTE + AENN	0.73	0.64	0.69	0.63	0.67
SVM	0.61	0.57	0.50	0.50	0.54
SVM + SMOTE	0.64	0.51	0.83	0.75	0.68
SVM + SMOTE + AENN	0.64	0.51	0.80	0.75	0.68
SVM + Feature Selection	0.70	0.64	0.55	0.50	0.6
SVM + Feature Selection + SMOTE	0.71	0.62	0.59	0.69	0.65
SVM + Feature Selection + SMOTE + AENN	0.70	0.58	0.79	0.62	0.67
Super Learner	0.82	0.77	0.52	0.69	0.70
Super Learner + SMOTE	0.73	0.67	0.60	0.62	0.66
Super Learner + SMOTE + AENN	0.72	0.66	0.62	0.62	0.65
Super Learner + Feature Selection	0.73	0.68	0.57	0.69	0.67
Super Learner + Feature Selection + SMOTE	0.74	0.67	0.61	0.71	0.68
Super Learner + Feature Selection + SMOTE + AENN	0.73	0.65	0.61	0.69	0.67

Table A3: *All AUC values for breast cancer data*

Model	Stage 1	Stage 2	Stage 3	Stage 4	Average
Random Forest	0.61	0.52	0.60	0.59	0.58
Random Forest + SMOTE	0.61	0.52	0.60	0.61	0.58
Random Forest + SMOTE + AENN	0.59	0.52	0.58	0.66	0.58
Random Forest + Feature Selection	0.59	0.52	0.57	0.60	0.57
Random Forest + Feature Selection + SMOTE	0.60	0.54	0.57	0.63	0.59
Random Forest + Feature Selection + SMOTE + AENN	0.60	0.56	0.60	0.65	0.60
Neural Network	0.57	0.50	0.55	0.50	0.53
Neural Network + SMOTE	0.59	0.50	0.54	0.61	0.56
Neural Network + SMOTE + AENN	0.58	0.50	0.53	0.62	0.56
Neural Network + Feature Selection	0.54	0.51	0.55	0.50	0.52
Neural Network + Feature Selection + SMOTE	0.57	0.52	0.55	0.63	0.56
Neural Network + Feature Selection + SMOTE + AENN	0.55	0.51	0.54	0.61	0.55
SVM	0.62	0.53	0.57	0.60	0.58
SVM + SMOTE	0.58	0.49	0.50	0.74	0.58
SVM + SMOTE + AENN	0.52	0.50	0.50	0.66	0.55
SVM + Feature Selection	0.59	0.53	0.56	0.55	0.56
SVM + Feature Selection + SMOTE	0.55	0.54	0.54	0.63	0.57
SVM + Feature Selection + SMOTE + AENN	0.59	0.56	0.52	0.64	0.58
Super Learner	0.60	0.52	0.59	0.50	0.55
Super Learner + SMOTE	0.61	0.52	0.60	0.60	0.58
Super Learner + SMOTE + AENN	0.58	0.51	0.50	0.67	0.57
Super Learner + Feature Selection	0.59	0.55	0.58	0.59	0.58
Super Learner + Feature Selection + SMOTE	0.61	0.54	0.58	0.64	0.59
Super Learner + Feature Selection + SMOTE + AENN	0.60	0.56	0.61	0.66	0.61

Table A4: *All AUC values for colorectal cancer data*

Model	Stage 1	Stage 2	Stage 3	Stage 4	Average
Random Forest	0.89	0.73	0.06	0.38	0.51
Random Forest + SMOTE	0.79	0.62	0.18	0.50	0.52
Random Forest + SMOTE + AENN	0.79	0.57	0.23	0.50	0.52
Random Forest + Feature Selection	0.80	0.62	0.14	0.39	0.49
Random Forest + Feature Selection + SMOTE	0.80	0.58	0.21	0.50	0.52
Random Forest + Feature Selection + SMOTE + AENN	0.80	0.54	0.23	0.50	0.52
Neural Network	0.87	0.62	0	0	0.37
Neural Network + SMOTE	0.79	0.44	0.27	0.54	0.51
Neural Network + SMOTE + AENN	0.82	0.32	0.43	0.30	0.47
Neural Network + Feature Selection	0.81	0.51	0	0	0.33
Neural Network + Feature Selection + SMOTE	0.76	0.50	0.41	0.35	0.51
Neural Network + Feature Selection + SMOTE + AENN	0.81	0.47	0.46	0.32	0.51
SVM	0.91	0.27	0	0	0.29
SVM + SMOTE	0.88	0.11	0.77	0.50	0.57
SVM + SMOTE + AENN	0.90	0.08	0.73	0.50	0.55
SVM + Feature Selection	0.81	0.52	0.12	0	0.36
SVM + Feature Selection + SMOTE	0.81	0.46	0.23	0.38	0.47
SVM + Feature Selection + SMOTE + AENN	0.84	0.29	0.69	0.25	0.52
Super Learner	0.88	0.73	0.04	0.38	0.51
Super Learner + SMOTE	0.79	0.60	0.21	0.25	0.46
Super Learner + SMOTE + AENN	0.78	0.57	0.26	0.25	0.47
Super Learner + Feature Selection	0.80	0.62	0.14	0.38	0.48
Super Learner + Feature Selection + SMOTE	0.80	0.57	0.25	0.43	0.51
Super Learner + Feature Selection + SMOTE + AENN	0.80	0.54	0.27	0.39	0.50

Table A5: *All Sensitivity values for breast cancer data*

Model	Stage 1	Stage 2	Stage 3	Stage 4	Average
Random Forest	0.52	0.19	0.58	0.19	0.37
Random Forest + SMOTE	0.53	0.26	0.49	0.23	0.38
Random Forest + SMOTE + AENN	0.61	0.32	0.28	0.33	0.39
Random Forest + Feature Selection	0.49	0.19	0.54	0.20	0.36
Random Forest + Feature Selection + SMOTE	0.52	0.28	0.46	0.29	0.39
Random Forest + Feature Selection + SMOTE + AENN	0.53	0.40	0.39	0.32	0.41
Neural Network	0.50	0.25	0.42	0	0.29
Neural Network + SMOTE	0.53	0.26	0.30	0.30	0.35
Neural Network + SMOTE + AENN	0.62	0.31	0.13	0.32	0.34
Neural Network + Feature Selection	0.54	0.17	0.42	0	0.28
Neural Network + Feature Selection + SMOTE	0.45	0.33	0.31	0.32	0.35
Neural Network + Feature Selection + SMOTE + AENN	0.49	0.36	0.21	0.27	0.33
SVM	0.51	0.17	0.60	0.21	0.37
SVM + SMOTE	0.89	0	0.01	0.64	0.39
SVM + SMOTE + AENN	0.99	0.01	0	0.32	0.33
SVM + Feature Selection	0.43	0.11	0.70	0.11	0.34
SVM + Feature Selection + SMOTE	0.54	0.31	0.26	0.32	0.36
SVM + Feature Selection + SMOTE + AENN	0.64	0.47	0.05	0.32	0.37
Super Learner	0.47	0.26	0.56	0	0.32
Super Learner + SMOTE	0.53	0.25	0.49	0.22	0.37
Super Learner + SMOTE + AENN	0.63	0.45	0	0.38	0.36
Super Learner + Feature Selection	0.42	0.35	0.50	0.20	0.37
Super Learner + Feature Selection + SMOTE	0.52	0.28	0.46	0.30	0.39
Super Learner + Feature Selection + SMOTE + AENN	0.54	0.40	0.39	0.34	0.41

Table A6: *All Sensitivity values for colorectal cancer data*

Model	Stage 1	Stage 2	Stage 3	Stage 4	Average
Random Forest	0.77	0.82	1	1	0.90
Random Forest + SMOTE	0.68	0.74	0.99	1	0.85
Random Forest + SMOTE + AENN	0.66	0.75	0.98	1	0.85
Random Forest + Feature Selection	0.68	0.75	0.99	1	0.85
Random Forest + Feature Selection + SMOTE	0.68	0.77	0.97	0.99	0.85
Random Forest + Feature Selection + SMOTE + AENN	0.67	0.77	0.96	1	0.85
Neural Network	0.63	0.81	1	1	0.86
Neural Network + SMOTE	0.60	0.78	0.97	0.97	0.83
Neural Network + SMOTE + AENN	0.52	0.75	0.92	1	0.80
Neural Network + Feature Selection	0.51	0.77	1	1	0.82
Neural Network + Feature Selection + SMOTE	0.65	0.77	0.94	0.99	0.84
Neural Network + Feature Selection + SMOTE + AENN	0.65	0.81	0.93	0.99	0.85
SVM	0.31	0.86	1	1	0.79
SVM + SMOTE	0.40	0.91	0.89	1	0.80
SVM + SMOTE + AENN	0.38	0.94	0.88	1	0.80
SVM + Feature Selection	0.60	0.75	0.99	1	0.84
SVM + Feature Selection + SMOTE	0.60	0.78	0.95	1	0.83
SVM + Feature Selection + SMOTE + AENN	0.57	0.87	0.88	1	0.83
Super Learner	0.77	0.81	1	1	0.89
Super Learner + SMOTE	0.68	0.74	0.99	1	0.85
Super Learner + SMOTE + AENN	0.66	0.74	0.98	1	0.85
Super Learner + Feature Selection	0.67	0.75	0.99	1	0.85
Super Learner + Feature Selection + SMOTE	0.67	0.77	0.97	1	0.85
Super Learner + Feature Selection + SMOTE + AENN	0.67	0.77	0.96	1	0.85

Table A7: *All Specificity values for breast cancer data*

Model	Stage 1	Stage 2	Stage 3	Stage 4	Average
Random Forest	0.71	0.84	0.62	0.99	0.79
Random Forest + SMOTE	0.68	0.78	0.71	0.99	0.79
Random Forest + SMOTE + AENN	0.56	0.71	0.88	0.99	0.79
Random Forest + Feature Selection	0.70	0.85	0.59	0.99	0.78
Random Forest + Feature Selection + SMOTE	0.69	0.79	0.70	0.98	0.79
Random Forest + Feature Selection + SMOTE + AENN	0.67	0.72	0.82	0.98	0.80
Neural Network	0.64	0.75	0.69	1	0.77
Neural Network + SMOTE	0.65	0.75	0.77	0.93	0.77
Neural Network + SMOTE + AENN	0.56	0.69	0.92	0.92	0.77
Neural Network + Feature Selection	0.53	0.86	0.68	1	0.77
Neural Network + Feature Selection + SMOTE	0.66	0.71	0.80	0.93	0.77
Neural Network + Feature Selection + SMOTE + AENN	0.61	0.65	0.87	0.94	0.77
SVM	0.73	0.90	0.55	0.98	0.79
SVM + SMOTE	0.27	0.98	1	0.83	0.77
SVM + SMOTE + AENN	0.06	0.98	1	0.99	0.76
SVM + Feature Selection	0.75	0.94	0.43	0.99	0.78
SVM + Feature Selection + SMOTE	0.55	0.77	0.83	0.94	0.77
SVM + Feature Selection + SMOTE + AENN	0.53	0.65	0.98	0.95	0.78
Super Learner	0.74	0.77	0.62	1	0.78
Super Learner + SMOTE	0.68	0.79	0.70	0.99	0.79
Super Learner + SMOTE + AENN	0.53	0.57	1	0.97	0.77
Super Learner + Feature Selection	0.76	0.74	0.66	0.99	0.79
Super Learner + Feature Selection + SMOTE	0.70	0.79	0.70	0.98	0.79
Super Learner + Feature Selection + SMOTE + AENN	0.67	0.72	0.82	0.98	0.80

Table A8: *All Specificity values for colorectal cancer data*

Rank	Stage 1	Stage 2	Stage 3	Stage 4
1	Selective or non-selective examination via percutaneous venous catheterization (4268)	Assessment of specimen breast tumor via operative session (4821)	PET whole body (5075)	PET whole body (5021)
2	Herceptin (4231)	Localization breast tumor (4653)	CT examination only prior to PET or SPECT (5041)	Sentinel node procedure (4978)
3	Iron (4152)	Mammography 3D (4641)	Sentinel node procedure (5005)	Carcinoma Antigen (4707)
4	Iron binding capacity (4138)	Age (4469)	MRI breast (4567)	Intravenous provision of bisphosphonates (4649)
5	Transferrin (4131)	Ultrasound breast (4452)	Regional lymph node extirpation (4495)	CT examination of the thorax, hart and large blood vessels (4635)
6	Ultrasound of the hart or thorax (3924)	Patient counseling during treatment with hormone therapy in non-metastatic tumors (4358)	Radiological examination of the shoulder, arm and hand (4425)	CT examination of the abdomen (4594)
7	Treatment contact social work (3908)	Mammography (4356)	Pathological examination of a needle biopsy or a complex cytological puncture (4420)	Patient support during hormone therapy treatment in metastatic or haematological tumors (4578)
8	Intake social work (3906)	Medical psychologist report (4185)	Erythocrytes (4334)	Carcinoembryonic antigen (4532)
9	Antibodies, IgT, IgG or IgA by immunoassay (3897)	Alkaline phosphatase (4093)	Physiotherapy session (4303)	Pathological investigation of simple biopsy or simple cytology (4493)

10	Non-clinical rehabilitation nursing (3774)	CT examination of the spine (4067)	Thromboplastin time (4238)	Removal of sentinel node (4321)

Table A9: *Scores of feature importance for the 10 highest scoring predictors in the breast cancer study without feature selection. The score values are shown between brackets. Medication prescription predictors are shown in red, diagnostic predictors are shown in green, invasive healthcare activity predictors are shown in blue and all other healthcare activity predictors are shown in black.*

Rank	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**
1	Call consultation to replace a repeat outpatient visit (3772)	Capecitabine (3884)	MRI rectum (4358)	Avastin 16ml (4240)
2	Treatment contact social work (3551)	Rectoscopy or proctoscopy (3713)	Diagnostic endoscopy of the colon (3953)	CT examination only prior to PET or SPECT (4171)
3	Intervention colonoscopy (3520)	Stomach resection (3519)	Pathological investigation of a complex resection (3918)	Chemotherapy by infusion or by injection in metastatic or hematological tumors (4115)
4	Endoscopic ileostomy (3469)	Oxaliplatin (3509)	Transferrin (3682)	Carcinoembryonic antigen (4071)
5	Therapeutic laparoscopy (3439)	Aersol treatment (3500)	Pathological examination of a simple large resection, moderately complex biopsy or special cytological preparation (3654)	Avastin 4ml (4046)
6	Differential count (3430)	SPECT of skeleton detail (3470)	Surgical removal of growths from subcutis (3647)	Complex molecular diagnostics (4010)

7	Chemotherapy by infusion or by injection in non-metastatic tumors (3413)	Mabthera 10ml (3449)	Limited CGA (3540)	Introducing a port-a-cath system (3908)
8	Case history and examination after referral for speech therapy (3413)	Mabthera 50ml (3437)	Insertion of central venous catheter (3525)	Endoscopic enterostomy (3852)
9	Construction of stoma (3382)	Ecalta (3392)	Removal of condition with the help of transanal endoscopic microsurgery (3475)	Urine screening (3647)
10	SPECT of skeleton detail (3340)	Static skeleton research (3381)	SPECT of skeleton detail (3465)	Rectoscopy or proctoscopy (3581)

Table A10: *Scores of feature importance for the 10 highest scoring predictors in the colorectal cancer study without feature selection. The score values are shown between brackets. Medication prescription predictors are shown in red, diagnostic predictors are shown in green, invasive healthcare activity predictors are shown in blue and all other healthcare activity predictors are shown in black.*

Rank	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**
1	Herceptin (517)	Age (590)	PET whole body (579)	Carcinoma Antigen (530)
2	Complete bone densitometric examination (512)	Assessment of specimen breast tumor via operative session (575)	Sentinel node procedure (561)	PET whole body (507)
3	Treatment wounds > 5 cm (494)	Patient counseling during treatment with hormone therapy in non-metastatic tumors (521)	MRI breast (533)	Carcinoembryonic antigen (499)

4	Chemotherapy by infusion or by injection in metastatic or hematological tumors (482)	Assessment radiological examination (503)	CT examination only prior to PET or SPECT (507)	CT examination of the abdomen (496)
5	Physiotherapy session (447)	Hospitalization (438)	Co-treatment outpatient clinic (485)	CT examination of the thorax, heart and large blood vessels (496)
6	Ultrasound of the heart or thorax (446)	Diagnostic puncture or biopsy of non-palpable abnormalities or organs, under MRI control (435)	Assessment of specimen breast tumor via operative session (458)	Sentinel node procedure (491)
7	HbA1c (438)	Bilirubin (430)	Removal of sentinel node (461)	Patient support during hormone therapy treatment in metastatic or haematological tumors (455)
8	CT examination of the thorax, heart and large blood vessels (437)	Treatment of large deep abscesses (422)	Physiotherapy session (422)	Pathological investigation of simple biopsy or simple cytology (437)
9	Pathological investigation of simple biopsy or simple cytology (436)	Removal of sentinel node (413)	Repeat outpatient visit (417)	Co-treatment outpatient clinic (424)
10	Patient support during hormone therapy treatment in metastatic or haematological tumors (435)	Duplex ultrasound (399)	Radiological examination of the shoulder, arm and hand (407)	Paclitaxel (395)

Table A11: *Scores of feature importance for the 10 highest scoring predictors in the breast cancer study with feature selection. The score values are shown between brackets. Medication prescription predictors are shown in red, diagnostic predictors are shown in green, invasive healthcare activity predictors are shown in blue and all other healthcare activity predictors are shown in black.*

Rank	Stage 1	Stage 2	Stage 3	Stage 4
1	Bilirubin (626)	Visit to emergency department (661)	MRI rectum (683)	Carcinoembryonic antigen (645)
2	First outpatient visit (591)	Teleconsult (579)	Endoscopic colon resection (672)	CT examination only prior to PET or SPECT (615)
3	Differential count (589)	Diagnostic endoscopy of the colon (555)	Diagnostic endoscopy of the colon (634)	Complex molecular diagnostics (612)
4	Carcinoembryonic antigen (566)	PET whole body (554)	Repeat outpatient visit (527)	Construction of stoma (550)
5	Complex molecular diagnostics (547)	Removal of condition with the help of transanal endoscopic microsurgery (534)	Prostate specific antigen (504)	Endoscopic enterostomy (546)
6	Construction of stoma (529)	Pathological investigation of a complex resection (522)	Enteroanastomosis surgery (500)	SPECT of skeleton detail (535)
7	Pathological investigation of a complex resection (494)	Capecitabine (520)	Therapeutic laparoscopy (489)	Prostate specific antigen (528)
8	Physiotherapy session (479)	CT examination of the thorax, heart and large blood vessels (513)	Teleconsult (488)	Mabthera 10ml (521)
9	Quantitative determination of an immunoglobulin (479)	SPECT of skeleton detail (499)	SPECT of skeleton detail (485)	Mabthera 50ml (509)

10	Radiological examination of the shoulder, arm and hand (407)	Intake social work (492)	Mabthera 10ml (471)	Capecitabine (507)

Table A12: *Scores of feature importance for the 10 highest scoring predictors in the colorectal cancer study with feature selection. The score values are shown between brackets. Medication prescription predictors are shown in red, diagnostic predictors are shown in green, invasive healthcare activity predictors are shown in blue and all other healthcare activity predictors are shown in black.*

Predictor	**Description**
Aerosol treatment	Provision of medication via mist -like gas.
Age	Age of the patient.
Alkaline phosphatase	Blood test to check for levels of alkaline phosphatases, which is a protein which could indicate a liver or bone disease.
Antibodies, IgT, IgG or IgA by immunoassay	Blood test to check for antibodies levels by imunnoassay. Specifically check for IgT, IgG or IgA levels, which are 3 different groups of immunoglobulins.
Assessment of specimen breast tumor via operative session	Operative removal of part of the breast tumor for assessment.
Assessment radiological examination	Assessment of any type of radiological examination.
Avastin	Medication consisting of antibodies. Specifically used in metastatic colorectal cancer.
Bilirubin	A blood test to check the levels of bilirubin, a breakdown product of hemoglobin.
Call consultation to replace a repeat outpatient visit	A call to the outpatient clinic to replace a visit to the outpatient clinic after one or more previous visits.
Capecitabine	Medication used for slowing down cancer growth.
Carcinoembryonic Antigen	Diagnostic test to determine expression levels the carcinoembryonic antigen. An over expression of this antigen indicates cancer presence.
Carcinoma Antigen	Diagnostic test to determine expression levels of the carcinoma antigen. This antigen is over expressed in a specific type of tumor called squamous cancer tumors.

Case history and examination after referral for speech therapy	Examination of the patient of his/her current state and medical history by a speech therapy practitioner.
Chemotherapy by infusion or by injection in metastatic or hematological tumors	Provision of chemotherapy specifically in metastatic or haematological tumors, and specifically via infusion or by injection.
Chemotherapy by infusion or by injection in non-metastatic tumors	Provision of chemotherapy specifically in non-metastatic tumors, and specifically via infusion or by injection.
Complete bone densitometric examination	Diagnostic scan for determining bone density.
Complex molecular diagnostics	Diagnostic tests performed on DNA and RNA.
Construction of stoma	Endoscopic construction of a stoma, which is an artificial anus.
Co-treatment outpatient clinic	Visit to the outpatient clinic
CT examination of the abdomen	CT scan for examination the abdomen.
CT examination of the spine	CT scan for examination of the spine.
CT examination of the thorax, hart and large blood vessels	CT scan for examination of the thorax, hart and large blood vessels.
CT examination only prior to PET or SPECT	CT scan for examination specifically prior to PET or SPECT scan.
Diagnostic endoscopy of the colon	A diagnostic examination of the colon via an endoscopy.
Diagnostic puncture or biopsy of non-palpable abnormalities or organs, under MRI control	A puncture or biopsy of any abnormality or organ for examination.

Differential count	A blood test to determine the percentage of each type of white blood cell present in the blood.
Duplex ultrasound	Examination of blood flow using ultrasound.
Ecalta	Medication prescribed for fungal infections in the blood or internal organs.
Endoscopic colon resection	Operative removal of the whole or part of the colon.
Endoscopic enterostomy	Operative procedure to divert the small bowel to the abdomen.
Endoscopic ileostomy	Operative procedure to divert the small bowel to the abdomen.
Entero-anastomosis surgery	Operative procedure to connect two parts of the bowel.
Entorostomy	Operative procedure to connect the stomach to the colon.
Erythrocytes	Blood transfusion.
First outpatient visit	The first visit to the outpatient clinic.
HbA1c	Examination of the average blood sugar level of the patient in the last few weeks.
Herceptin	Medication for immunotherapy. This medicine is used specifically in HER2-positive tumors, to prevent the growth of the tumor and to prevent the tumor from spreading.
Hospitalization	Hospitalization of the patient.
Insertion of central venous catheter	Insertion of a catheter in the central vein.
Intake Social Work	Intake meeting with a social worker from the hospital
Intervention colonoscopy	Any assistence during a colonoscopy (e.g. treatment for bleeding).
Intravenous provision of bisphosphonates	Provision of bisphosphonates in the vein of a patient. Bisphosphonates are drugs that prevent the loss of bone density.
Introducing a port-a-cath system	Insertion of a port-a-cath system, which is a catheter in the central vein.
Iron	A blood test to check for iron levels in a patient's blood.
Iron bonding capacity	A blood test to see if a patient has too little or too iron in his or her blood.
Limited CGA	Assessment of patient status for frail or older patients.
Localization of breast tumor	Mammography for localizing breast tumor.
Mabthera	Medication containing antibodies, used in specific types of cancer.
Mammography	General mammography for breast examination.
Mammography 3D	Mammography in 3D for breast examination.

Medical psychologist report	Report by a medical psychologist.
MRI breast	MRI scan of the breast for examination.
MRI rectum	MRI scan of the rectum for examination.
Non-clinical rehabilitation nursing	Rehabilitation activity without any clinical process. Clinical processes include diagnosing or treatment of a patient.
Oxaliplatin	Medication used in treatment of colorectal cancer.
Paclitaxel	Medication for chemotherapy. This medicine is used to prevent cell division in tumors, and is used both in metastatic and non-metastatic cancers.
Pathological investigation of simple biopsy or simple cytology	Pathological investigation of a small sample of tissue from a patient's body.
Pathological investigation of a complex resection	Pathological inverstifation of a sample tissue from a patient's body after a complex resection.
Pathological examination of a needle biopsy or a complex cytological puncture	Pathological examination of a sample tissue from a patient's body acquired from a needle biopsy or a complex cytological puncture.
Pathological examination of a simple large resection, moderately complex biopsy or special cytological preparation	Pathological examination of a sample tissue from a patient's body acquired from either a simple large resection, moderately complex biopsy or a special cytological preparation.
Patient counseling during treatment with hormone therapy in non-metastatic tumors	Counseling during any type of hormone therapy specifically for patients with either metastatic or haematological cancer.
Patient support during hormone therapy treatment in metastatic or haematological tumors	Support during any type of hormone therapy specifically for patients with either metastatic or haematological cancer.
PET whole body	PET scan of the whole body for examination.
Quantitative determination of an immunoglobulin	A blood test to determine the levels of an immunoglobulin.

Physiotherapy session	A session for physiotherapy treatment.
Prostate specific antigen	Blood test to check for prostate specific antigen levels in the patients blood.
Radiological examination of the shoulder, arm and hand	A radiological examination looking at the entire shoulder, arm and hand.
Radiological examination of the skull	A radiological examination looking at the skull.
Rectoscopy or proctoscopy	An internal examination of the anus and rectum.
Regional lymph node extirpation	Operative removal of one or more lymph nodes.
Removal of condition with the help of transanal endoscopic microsurgery	Removal of any condition via endoscopic microsurgery.
Removal of sentinel node	Operative removal of the sentinel node.
Repeat outpatient visit	Visit to outpatient clinic after a previous visit.
Selective or non-selective examination via percutaneous venous catheterization	Diagnostic examination by inserting a catheter in a peripheral vein.
Sentinel node procedure	Removal of some of the sentinel nodes for examination.
SPECT of skeleton detail	SPECT scan of the skeleton for examination.
Static skeleton research	Examination of a patient skeleton in a static position.
Stomach resection	Removal of all or a part of the stomach.
Surgical removal of growths from subcutis	Operative procedure to remove any growths from the subcutis, which is one of the lower layers of skin.
Teleconsult	Online consultation.
Therapeutic laparoscopy	Operative procedure in the abdominal wall.
Thromboplastin time	A screening test that helps evaluate a patient's ability to appropriately form blood clots.
Transferrin	A blood test to check for transferrin levels. Transferrins are iron-binding blood plasma glycoproteins.

Treatment contact social work	Follow-up appointment with a social worker from the hospital.
Treatment of large deep abscesses	The removal of any large deep abscesses.
Treatment of wounds > 5 cm	Treatment of any wound larger than 5 cm, with or without wound edge excision. This includes examination, cleaning and bonding/glueing.
Ultrasound breast	Ultrasound imaging of the breast for examination.
Ultrasound of the hart or thorax	Ultrasound imaging of the hart or thorax for examination.
Urine screening	Examination of multiple substance levels in the urine of a patient.
Visit to emergency department	A visit to the emergency department for first aid help.

Table A13: *Descriptions of the predictors included in the top 10 most important predictors for any of the models*