Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 17 (2013) 789 – 796

Procedia
Computer Science

Forecasting trends in the Healthcare sector.

Ruben Lam

VU University Amsterdam, Master Business Analytics
De Boelelaan 1081, 1081HV, Amsterdam, the Netherlands

Abstract

Modern healthcare has become a trending topic nowadays. The most important aspect in healthcare is improving the health of patients. The objective of this paper is to create a predictive risk model using factors such as the amount of consults per patient, the medication per patient and the age of the patient. The model should be able to predict the risk of contracting colorectal cancer. The technique used for this is making a risk model, especially the use of the C5.0 model. Understanding the data and data preparation are the first important steps in the process. This will result in a modelling set. The next step is to model this set, taking different machine learning techniques into account. In addition, different configurations for model options and variable subsets are applied.
In this paper, the most optimal C5.0 decision tree with temporary variables gave us an accuracy of 84,3% for the total population. The model is applied to predict CRC for patients in Utrecht. However, this model seems to be applicable to large datasets and for various analyses. The use of predictive models improves the quality of the healthcare section. Therefore, there is an urgent need for detailed research into the use of predictive modelling.

1. Introduction

In recent years, researchers have become increasingly interested in forecasting trends in diseases in healthcare. In general, when patients show the same symptoms, they are more likely to have the same disease. Colorectal cancer (CRC) is the third most common cancer in men and the second most common cancer in women worldwide. More than 1.2 million new cases of colorectal cancers are diagnosed globally, with more than 600.000 related deaths in 2008 (Jemal, Siegel, Ma, & Zou, 2014). Survival rates are directly related to the time of diagnoses. Therefore, is it important to detect CRC at the most early stage. This early detection can be achieved by using many machine learning techniques. Operation research projects and data mining techniques are applied in the healthcare sector every day to achieve these life savings.

Previous studies have reported on predicting cancer. It was pointed out by Bellaachia and Guven (Bellaachia & Guven, 2006) that breast cancer survivability can be predicted used data mining techniques. They investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network and the C4.5 algorithm. They showed that the C4.5 algorithm gave the best performance, when predicting the survival rates of breast cancer. To date, several studies have examined breast cancer and other common cancers. Only few have examined predicting colorectal cancer. This is the main reason to discuss forecasting colorectal cancer.

This study describes an improved risk model from the model about predicting CRC, made last year (Hoogendoorn, Moons, Numans, & Sips, 2014). This paper describes how to generate "a predictive model using the CHAID decision tree learner based on anonymously extracted Electronic Medical Records:". The data, mentioned in section 2, is similar and a decision tree is used in both projects. In contrast, the used specific model and input for the model are different. Furthermore, for the input of the model we use temporary data mining. This means that we split all patient data into small time sections and use those time sections as input for our model.

Forecasting can be done by the use of data mining, especially the use of modelling time series. Few reports have discussed this subject. Data mining is the technique for discovering patterns in large data sets. Several methods like data pre-processing and modelling are used to transform data into understandable conclusions. Modelling involves tasks like clustering, classification, regression and forecasting. Data mining methods are used to analyse large data sets. In this case, we analyse a patient dataset from general practitioners in Utrecht, the Netherlands.

The objective of the present study is to create a predictive risk model using several factors, including time dependent variables, described in this paper. The model should be able to predict the risk of contracting colorectal cancer. Furthermore, we determine whether the condition and the medication of this patient affect this risk. There are several steps in this process: pre-processing and understanding the data, making a risk model and giving statistical correct conclusions and outcomes. Section 2 will give all the crucial information for the dataset. The rest of this paper is structured as follows. Section 3 gives an explanation about the used methods . Section 4 report the experimental results and the conclusion. We end this paper with a discussion.

2. Dataset Description

The data originates from all general practitioners (GP) from Utrecht, Netherlands. This medical data consist of three parts. Information about the patients, the medication, the consults and about referral to medical specialists. The first part is about the general information of the patients. Each patient is identified by their year of birth, gender, partial ZIP-code, practice code and the current period they are registered at the general practitioner. In total the dataset consist of 142,061 patients. Further, the data covers a period between January 1, 2006 and December 31, 2011. For this research we choose a period of one year that a patient was active at the GP. This year is taken random within the active period for patients that are not diagnosed with CRC. For patients diagnosed with CRC we choose the year before CRC is first diagnosed. Accordingly, the data has been anonymized by using random patient identification numbers instead of full names.

The second part of the data concerns the consults. For every consult the diagnose is recorded with an ICPC code. D75 is the ICPC code for CRC. One or more consults are recorded for every patient. To illustrate this point figure 1 and figure 2 show the distribution of consults for a year of data. The average amount of consults per patients is 2,3. Figure 2 shows only patients with CRC. In contrast, the average amount of consults per CRC-patient is 2,7. This dataset consist of 1,543,670 consults. When we select only patients that are a year or longer active at the GP, this reduces to 242,796. After selecting these patients we now have 103,372 patients, 477 with CRC diagnosed.

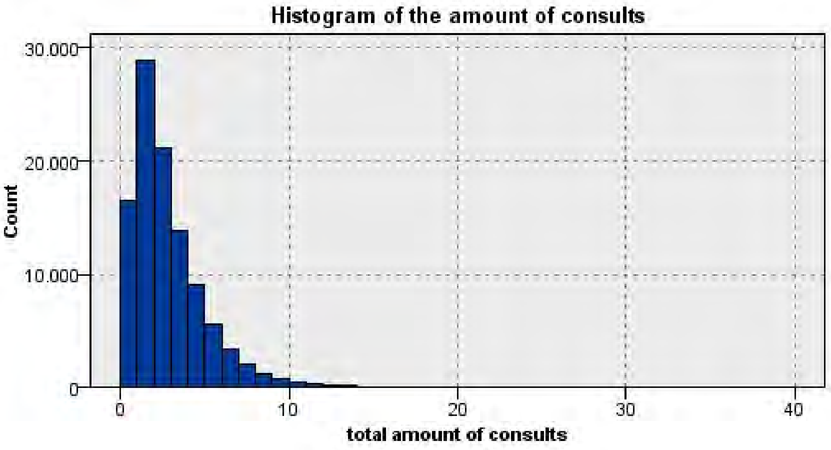

Figure 1: Histogram of the amount of consults in a selected year of data

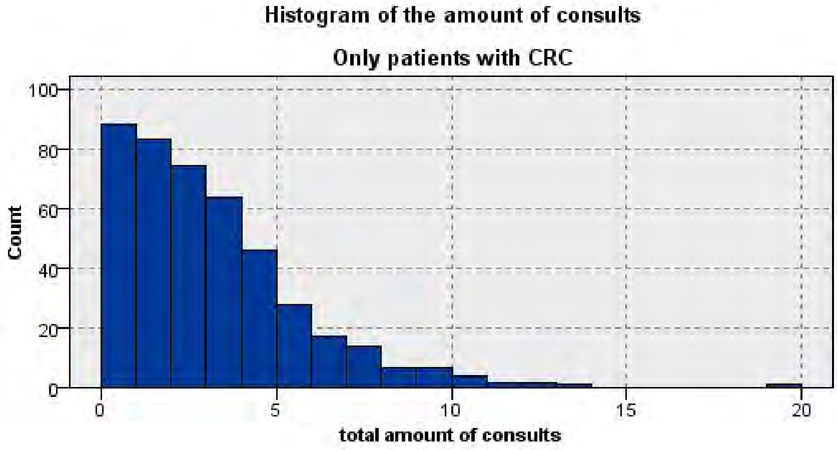

Figure 2: Histogram of the amount of consults for CRC patients in a selected year of data

The third part of the data gives information about the medication of every patient. Every medication is recorded with an ATC code. ATC stands for Anatomical Therapeutic Chemical. Other attributes are the period of prescription, route of administration and patient number. There are 91 different ATC codes. Figure 3 shows the distribution of ATC codes. It makes clear that there is a large variety in medication. The most given medication is code C, this is medication related to the cardiovascular system.

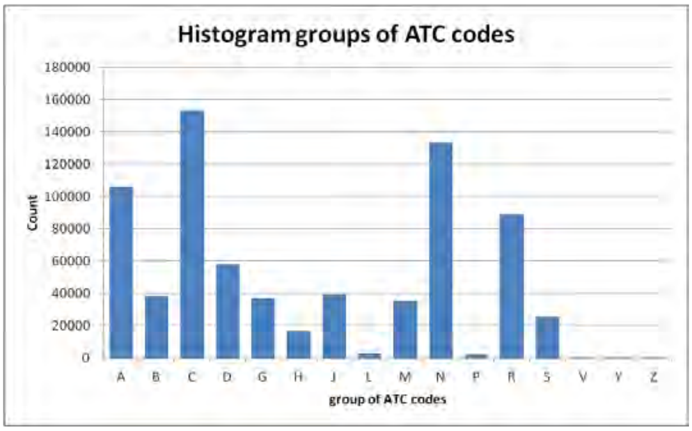

Figure 3: Histogram of all groups of ATC-codes

Referral to medical specialists is the last set of data. When a patient at the GP is referred to a medical specialist at the hospital or clinic it is stored in this data. Every record consist of a patient number, date of referral and the specialist.

For every patient, a year of data is chosen that he or she was registered at the GP. When a patient is diagnosed with CRC, we choose the year before the diagnose is confirmed. The distribution of the different parts of data is shown in figure 4, figure 5 and figure 6. Looking at the consults, it is clearly visible that the amount of consults is higher in the fourth quarter of the active year at the GP. In particular for patients with CRC. This is what we expected, because this is the last quarter before he or she is diagnosed with CRC. In addition, the amount of medication for CRC patients in quarter four is higher than other quarters. This is tested with a non-paired students t-test resulting in all p-values lower than 0.05. Figure 5 clearly shows the difference in medications given.

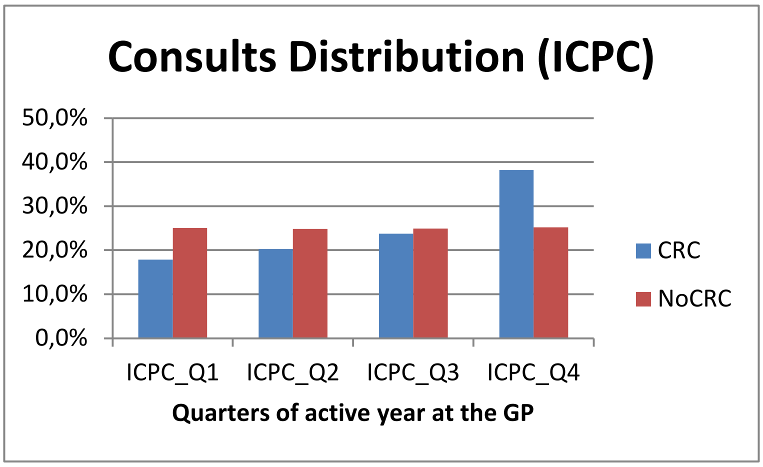

Figure 4: Distribution per quarter for consults

There is a clear difference for quarter four, the quarter before diagnosed with CRC, for consults and medication. The non-paired students t-test shows, when testing the alternative hypothesis: “true difference in means is not equal to 0”, a p-value lower than 0.05 for consults and medication. In contrast, the students t-test shows a p-value of 0.14 for the referrals in the fourth quarter.

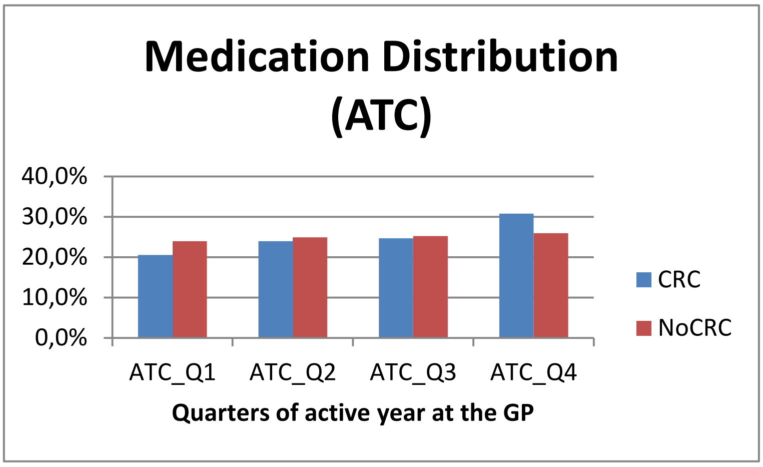

Figure 5: Distribution per quarter for medication

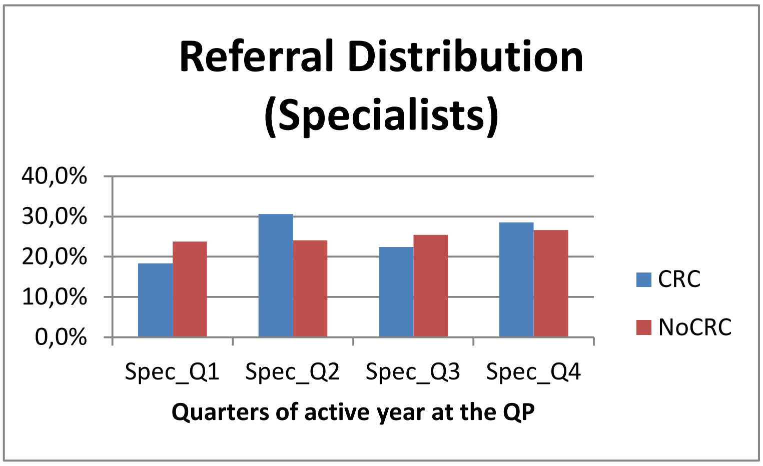

Figure 6: Distribution per quarter for referrals

Looking at the complete dataset, we can look at some other statistics. The distribution of men (45.289) and women (58.083) should be almost equal, because we don't model on the gender. Figure 7 shows that the dataset consist of slightly more women. We assume that this will not affect our model.

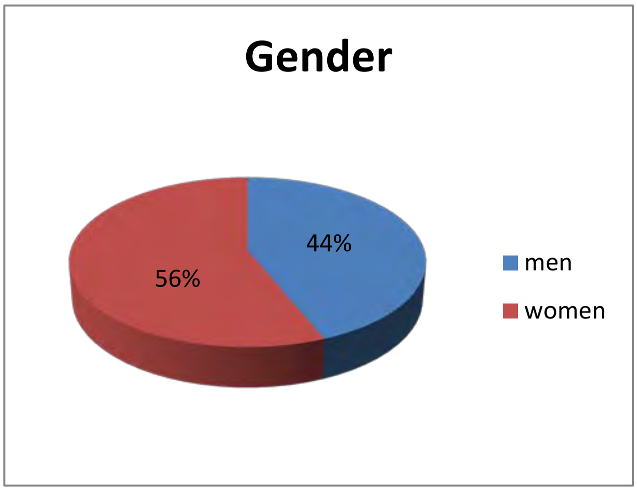

Figure 7: The distribution of men and women in the data

3. Method: Data mining

3.1 Dataset preparation.

Generally, Every dataset needs preparation for modelling. It is important to realise what input the model needs. As mentioned before in chapter 2 we choose a period of one year for every patient. When looking at trends in the data, we need a time component. Therefore we split this year in four periods, quarters of a year. The data consists of a date for every consult. We count these consults for every quarter. We achieve this result by selecting only a single quarter of the data and then count the records per patient.

The next step is to split this count per ICPC to determine the diagnose of every consult. The data consist of 714 ICPC codes. When we split on these codes, the data gets to sparse. However, we can split the data on groups of ICPC codes. This means that we combine all ICPC codes A01...A99 to code A. We do this for every ICPC code, except D75 since this is CRC. For example, ICPC code D are all diagnoses with digestive concerns and ICPC code S for sensory organs. Figure 4 shows an example of our current result per patient. In this example July 30 is the date CRC was diagnosed. For patients that are not diagnosed with CRC we take a random year within their period active at the GP.

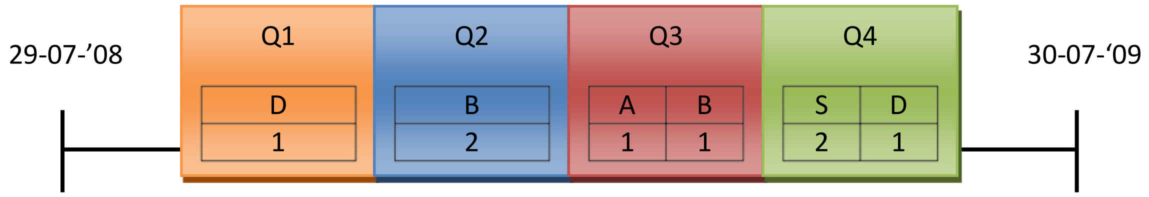

Figure 8: Visualization of the ICPC codes of one patient for one year

The data preparation of this data, as we see in figure 8, can be done for other parts of the data. Referral to different medical specialists and the medication of every patient can be counted as well. For medication the same technique as consults is applied. For every ATC code we split on groups of ATC codes. It should be noted that ATC group codes do not relate to ICPC group codes. In other words, medication code A does not relate to consults code A.

In the distribution of our data from section 2 we see an overall increasing trend towards the fourth quarters. This trend can be measured for every patient using linear regression. For example, a patient has three, four, one and five consults in every quarter. We calculate a trendline for this patient and the scope, which will be positive for most patients. In our example the trend is 0.3. For every group of ICPC, Spec and ATC codes, slopes are added as attributes to our modelling set. In addition, we add a slope for the sum of ICPC, Spec and ATC codes.

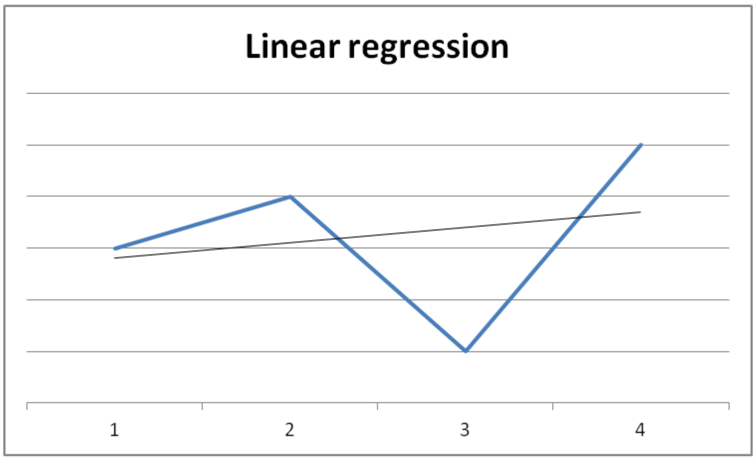

Figure 9: Visualisation of linear regression

The last step in our data preparation is adding the totals of all quarters for every part of the data. For example, patient 2152 was given 4 prescriptions in total for quarter two. Adding these attributes to our preparation gives 489 available variables in total to model this data. Table 1 shows all the variables available in the modelling set. Every record represents a single patient.

Variables	Datatype	Explanation
Age	Integer	The age per patient at the end of the year active at the GP.
ICPC_A_Q1 - ICPC_Z_Q4	Integer	The amount of consults per quarter per ICPC code on high level.
Spec_acupunctuur_Q1 - Spec_verloskunde_Q4	Integer	The amount of referrals per quarter.
ATC_A_Q1 - ATC_Z_Q4	Integer	The amount of medication per quarter per ATC code on high level.
ICPC_Q1..Q4_SUM, Spec_Q1..Q4_SUM, ATC_Q1..Q4_SUM	Integer	The total amount of ICPC, ATC and referral per quarter
ICPC_A_Slope - ICPC_Z_Slope	Double	The slopes for groups of ICPC codes. These can be negative or positive.
Spec_acupunctuur_Slope - Spec_verloskunde_Slope	Double	The slopes for groups of Spec codes. These can be negative or positive.
ATC_A_Slope - ATC_Z_Slope	Double	The slopes for groups of ATC codes. These can be negative or positive.
ICPC_Slope, Spec_Slope, ATC_Slope	Double	The slopes for ICPC, ATC and referral. These can be negative or positive.
Total_ICPC_Spec_ATC	Integer	The total amount of consults, medication and referrals per patient.

Table 1: Variables available in the modeling set

3.2 The algorithm and setup

During this stage of the investigation, many models have been tested. Certain models did not fit the data, because methods like neural networks, Bayesian networks and Association rules do not give appropriate results. A decision tree is appropriate because the algorithm is relatively simple to explain and it fits the data.

The C5.0 algorithm is chosen to model this data. This algorithm is developed by Ross Quinlan, a computer science researcher in data mining and decision theory. C5.0 uses less memory and is faster than C4.5, an earlier developed version of this algorithm. More information about C5.0 is given by Barzdins (Barzdins, Gosko, Rituma, & Paikens, 2014). However, the specific explanation and functions are not given by Quinlan. Therefore we will explain C4.5 to give a brief overview of the algorithm.

C4.5 is a decision tree builder from a set of training data, using the concept or information entropy (Quinlan, 1993). Entropy is a measure of unpredictability of information content. We define the classes $\{C_1, C_2, \dots, C_k\}$. The training data is a any set $S = s1, s2, \dots$ of already classified samples. Each sample s_i consists of a multi dimensional vector with attributes or variables. In addition, every sample contains the class in which s_i falls.

The tree makes a decision on every node. It chooses the attribute of the data that most effectively splits its set of samples into subsets. The normalized information gain is the criteria used for splitting. The variable that has the highest information gain is mainly used for the decision in each node. This is a recursive process, the algorithm will go one step down in the tree and continue on smaller subsets.

Entropy and information gain

We define $f(C_i, S)$ as the number of samples that belong to a certain class. In addition, we define $|S|$ as the number of samples in set S. Then the entropy is defined as (Brissaud, 2005):

$$Entropy(S) = -\sum_{i=1}^{k}\left[\left(\frac{f(C_i,S)}{|S|}\right) * \log_2\left(\frac{f(C_i,S)}{|S|}\right)\right]$$

We define the information gain for the training set as follows:

$$Entropy_x(T) = \sum_{i=1}^{n}\left[\left(\frac{|T_i|}{|T|}\right) * Info(T_i)\right]$$

The gain is then defined as the difference in entropy: $Gain(X) = Entropy(S) - Entropy_x(T)$
After the gain is calculated, the attribute with the highest information gain is chosen.

This algorithm has a few base cases. The base case determines the termination condition for recursion. First, in a node, it can appear that every sample already belongs to the same class. When this occurs, it simply chooses this class. Secondly ,when none of the attributes provides any information gain, the C4.5 algorithm will take the expected value of classes higher in the tree.

In pseudocode, the algorithm for building decision trees, according to Kotsiantis et al. (Kotsiantis, Zaharakis, & Pintelas, 2007):

```
1. Check for base cases
2. For each attribute x
       o Find the normalized information gain ratio from splitting on x
3. Let x_best be the attribute with the highest normalized information gain
4. Create a decision node that splits on x_best
5. Recurse on the subsets obtained by splitting on x_best, and add those nodes as children of
   the node
```

Model options

This section will report on several techniques and options that can be applied when using a decision tree learner model. The result section will discuss which techniques are used for the final predictive model.

Feature selection can be applied to reduce the attributes, creating a more manageable set of variables for modelling. This will simplify and narrow the scope of the features, which is essential for our predictive model (Liu, Li, & Wong, 2002). The predicted target is set to CRC, which is a true/false parameter. Different predictors are used, such as Pearson's Chi-square, Cramer's V and the Likelihood Ratio Chi-square. These predictors are sorted by their p-value per attribute or set of attributes towards our target. Table 4 gives an example of the results for feature selection. In common, when a variable has an importance value lower than 0.9, it's not used in the model. The final selection of variable's for the best result will be shown in the result section.

Variable	Importance value	Classification
Age	1.0	Important
ATC_Code_B	1.0	Important
Specialism_internal_medicine	0.997	Marginal
...	...	...

Table 2: The results of a feature selection example

Cross-validation can be applied to C5.0. This is a model validation technique to determine how the intermediate results of an analysis will generalize. We can estimate the accuracy for our predictive model with cross-validation and improve our model. For this dataset a 10-fold cross-validation is chosen, since this has been proven to give a reasonably good estimate (Kohavi, 1995). This technique splits the data in ten equally sized samples, a single part for testing and nine parts to test the model. We repeat this validation 10 times, for every sample. Figure 10 illustrates a 5-fold cross-validation.

The C5.0 algorithm makes use of information gain when calculating the optimal classification. This is the expected information that you will obtain when going from a prior state to the next state. Many aspects of information gain are discussed by Nowozin (2012).

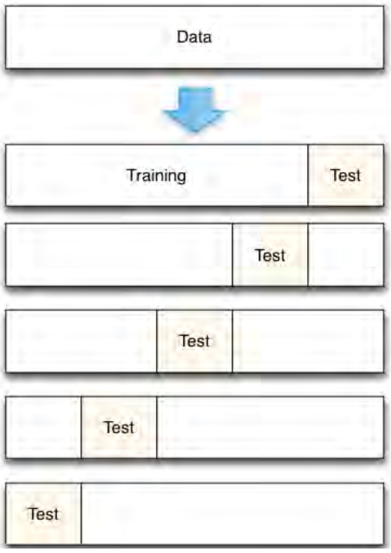

Figure 10: 5-fold cross validation

Several Options can be managed when applying the algorithm. We apply misclassification costs, because the dataset is highly unbalanced (477 CRC patients and 103,372 patients in total). We apply the ratio between CRC patient and a non-CRC patient as misclassification costs. The C5.0 algorithm takes this ratio into account to construct the optimal model. Furthermore, cross-validation is set to 10 folds as mentioned before and we focus on accuracy.

Boosting

This machine learning technique is an algorithm to reduce bias. There are many different boosting algorithms available. Boosting improves weak classifiers with respect to the distribution. In other words, boosting makes a single strong classifier from many weak classifiers. This process is iterative and the data is reweighted after every classifier. AdaBoost (Freund, Schapire, & Abe, 1999), will even try to adapt the weak learners to optimize the learning process.

4. Results

4.1 Variable selection.

Our modelling set contains of 489 variables. To prevent the model from overfitting, a selection of different sets is made. Many options are tested and the best five variable sets are chosen, option 1 to option 5. In addition, we added the model from Hoogendoorn M et al. as option 0, to compare our results. Note that we run that model with C5.0, in contrast to the Chaid decision tree. The counts per quarter and per code are always used, since they temporal effects are the main attributes of our model. Table 3 shows the five selected sets of variables and option 0.

Variables	Option 0	Option 1	Option 2	Option 3	Option 4	Option 5
Non-temporal variables						
Age	x			x	x	x
ICPC_A01 - ICPC_Z67	x					
ATC_A01 – ATC_V09	x					
Spec_cardiology – Spec_urology	x					
ICPC_Group_A – Z	x					
ATC_Group_A – V	x					
Temporal variables						
ICPC_A_Q1 - ICPC_Z_Q4		x	x	x	x	x
Spec_accupunctuur_Q1 - Spec_verloskunde_Q4		x	x	x	x	x
ATC_A_Q1 - ATC_V_Q4		x	x	x	x	x
ICPC_Q1..Q4_SUM, Spec_Q1..Q4_SUM, ATC_Q1..Q4_SUM			x		x	x
ICPC_A_Slope - ICPC_Z_Slope						x
Spec_acupunctuur_Slope - Spec_verloskunde_Slope						x
ATC_A_Slope - ATC_Z_Slope						x
ICPC_Slope, Spec_Slope, ATC_Slope			x		x	x
Totals						
Total_ICPC_Spec_ATC	x		x		x	x

Table 3: The five variable selections

For the first run, without any model options, we do not use feature selection. Boosting is not applied and the C5.0 mode will be set to default. Nevertheless we do use 10-fold cross-validation, because this gives improved results without increasing the runtime. To validate the outcome of our results we report how many predictions of CRC are true. Therefore, we add the true negative and the true positive and calculate the percentage true of the total amount of patients. We call this the hitrate and the results are shown in table 4.

	Option 0	**Option 1**	**Option 2**	**Option 3**	**Option 4**	**Option 5**
Hitrate	79,2%	72,3%	69,8%	73,6%	73,4%	74,9%

Table 4: The hitrate for every variable selection

A confusion matrix is a matrix that shows the performance of an algorithm, using the actual and the predicted outcome. The confusion matrix for option 5, the best result with temporary variables, is shown in table 5. The rows show the actual value of having CRC, the columns the predicted value. Note that the total count adds up to about 41.000 patients, since we use a test set of 40% of the data.

Confusion Matrix		**Predicted**	
		No-CRC	CRC
Actual	No-CRC	30.844	10.322
	CRC	27	150

Table 5: The confusion matrix for variable option 5.

The conclusion of table 5 is that we have a lot of patients that don't have CRC and are predicted having CRC. These patients in the model have a higher risk of getting CRC than No-CRC patients.

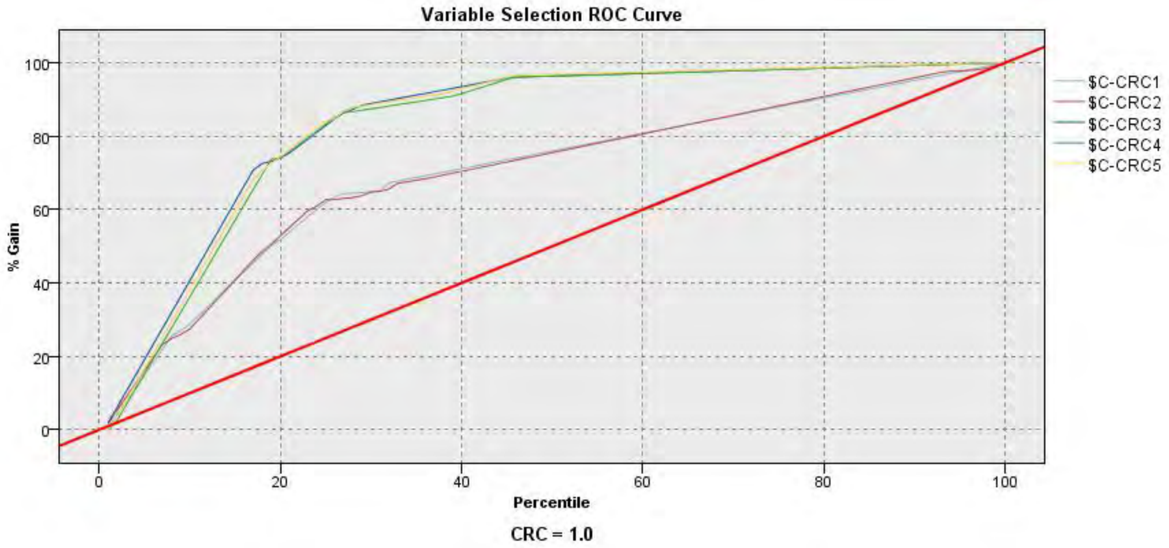

Figure 11: The ROC curve for every variable selection

Conclusion

To illustrate the information gain we create a receiver operating characteristic curve. These curve is called the ROC curve and shows the performance of a binary classifier. By plotting the fraction of true positives of the total actual positives at the vertical axis and the fraction of false positives at the horizontal axis, a ROC curve is created. This provides us with an analysis to see the model that is most optimal. As clearly visible in figure 11, option 3, option 4 and option 5 have the most area under the curve and are therefore most optimal. There is a clear significance between these options and option 0, 1 and 2. The non-paired students t-test shows, when testing the alternative hypothesis: "true difference in means is not equal to 0", a p-value higher than 0.9. Comparing, for example, option 1 and option 2 the t-test shows a p-value of 0.9959, meaning the ROC curves do not seem to be different, which makes sense. In addition, option 3 for example, has an area under the curve (AUC) of 0.803, while option 1 only has an AUC of 0.660. We will elaborate on the ROC curve of option 0 in the conclusion of section 4.2.

Important predictors of the best model with temporary variables

First, we look at the ICPC codes. The most important predictors, based on information gain, are group Y in Q1, group H in Q4 and group D in Q3. Remember that Q4 means the fourth quarter active at the GP and the quarter before first diagnosed with CRC, only for CRC patients of course. Second, we look at ATC codes. The most important groups here are D in Q1, J in Q3 and M in Q3. Finally, for referrals to specialism only physiotherapy in Q2 is classified as an important predictor. Thus we see that the temporal variable's are important in predicting the target value, CRC. The association of these ICPC codes with CRC has to be studied.

4.2 Model variation

After choosing four different sets of variables to model on, it is important to fine tune the model. Therefore, we experiment with C5.0, which has a few options to change the outcome of the model. Running different models will show the most optimal model parameters. This is called trial and error.

First, we choose to set feature selection on. This means that we select a subset of variables. This technique allows us to identify the most important variables to be used in our model. Removing un important variables can help the model improve. In addition, the building of the model will speed up and it will be easier to deploy the model, because simpler models with fewer input are more practical. Table 8 shows the outcome of feature selection at the second row with results.

Second, we set the minimum records per child branch to two and the pruning severity to 75. Pruning severity has shown effective on growth (Kumar, Katiyar, Singh, & Rajkumar, 2014). Similar to feature selection, the result is slightly improved and shown in table 8.

Finally, the boosting option is used. The machine learning technique boosting is explained in section 3.2. For our model it significant increases our runtime, but the results are improved as well. The parameter trials is set to ten and twenty, to see whether that would improve the model even more. Yet, the outcome are worse. Table 6 shows the results of boosting in the last two rows.

Hitrate	**Model Options**			**Variable sets**					
Description	**Feature selection**	**C5.0 mode**	**Boosing trials.**	**Option 0**	**Option 1**	**Option 2**	**Option 3**	**Option 4**	**Option 5**
First run, without any model options. (Section 4.1)	Off	Default	0	79,4%	72,3%	69,8%	73,4%	73,6%	74,9%
Feature selection on	On	Default	0	85,1%	70,4%	71,3%	73,1%	75,2%	73,9%
C5.0 adjustments	On	Adjusted	0	84,8%	71,0%	72,2%	73,1%	76,1%	72,0%
Boosting on with 10 trials	On	Adjusted	10	92,2%	80,9%	77,4%	80,9%	84,3%	78,6%
Boosting on with 20 trials	On	Adjusted	20	91,1%	76,4%	77,4%	79,4%	80,6%	83,4%

Table 6: The hitrate, outcome, of the variable sets with different model options.

The confusion matrix for option 4 with an hitrate of 84,3%, the best result with temporary variables, is shown in table 7. The rows show the actual value of having CRC, the columns the predicted value. Note that the total count adds up to about 41.000 patients, since we use a test set of 40% of the data.

Confusion Matrix		**Predicted**	
		No-CRC	CRC
Actual	No-CRC	34.765	6.401
	CRC	58	119

Table 7: The confusion matrix for variable option 4 with modeling options according to table 6

The conclusion of table 7 is same as the conclusion of table 5. There are a lot of patients that don't have CRC and are predicted having CRC. These patients have a higher risk in the model of getting CRC than No-CRC patients.

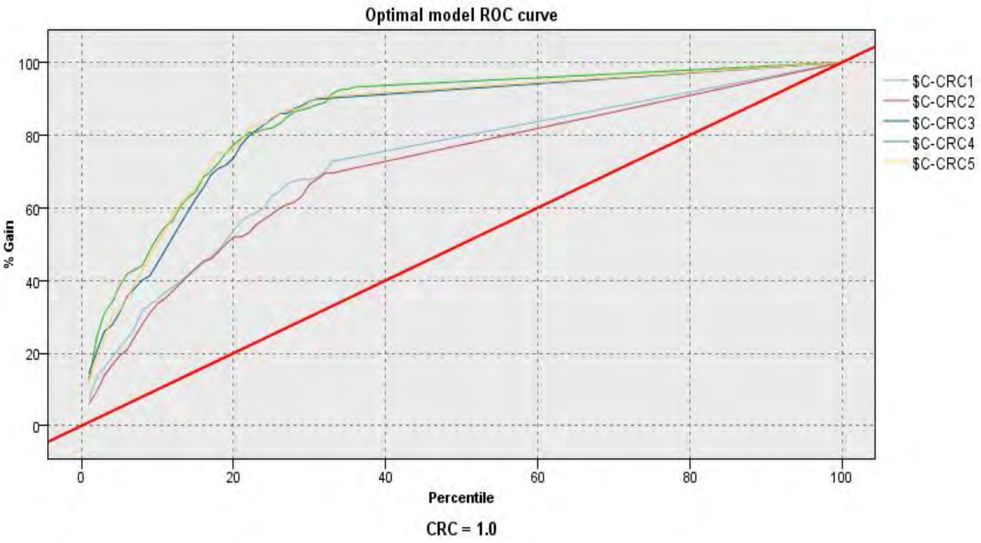

Figure 12: The ROC curve for every variable selection with the best model options

Finally, we calculate the ROC curves of the risk model, before adding all temporary variables (Hoogendoorn, Moons, Numans, & Sips, 2014) . Figure 13 shows the results, where it is again clearly visible that extra modelling options give better results. In the conclusion we compare the AUC values, area under the curve, with the current results.

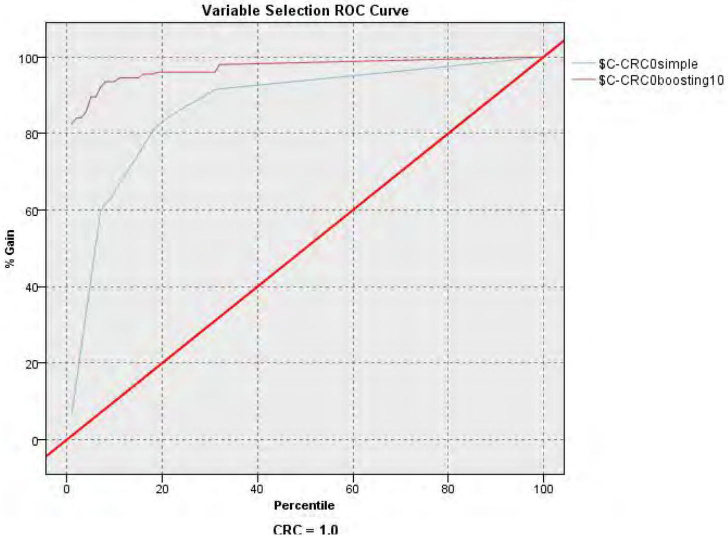

Figure 13: The ROC curves without temporary variables

Conclusion

As clearly visible in figure 12, option 3, option 4 and option 5 have the most area under the curve and are therefore most optimal. There is a clear significance between these options and option 1 and 2. In accordance with section 4.1, the non-paired students t-test shows that the difference between the low and high curves is significant.
Boosting with 10 trials, Feature selection on, the minimum records per child branch to two and the pruning severity to 75 gave the best results for five of the six variable selections. Figure 12 shows the ROC curves for these model parameters. In comparison to figure 11, the area under the curve has slightly increased. This is not significant according to the students t-test, tested with the same hypothesis as mentioned in the conclusion of section 4.1. Thus the information gain might be higher after using many modelling options. However, the hitrate does increase with more than 10%.
Table 8 shows the results when comparing all AUC values. Note that option 0 is without temporary variables (Hoogendoorn, Moons, Numans, & Sips, 2014).

The variable set	AUC without any modelling options	AUC with feature selection, adjusted model options and boosting on.
Option 0	0.806	0.898
Option 1	0.660	0.687
Option 2	0.669	0.679
Option 3	0.803	0.801
Option 4	0.809	0.825
Option 5	0.817	0.796

Table 8: The AUC values for all ROC curves

Final model with temporary variables
Our final model with temporary variables, is model option 4 with 10 trial boosting applied. This model showed a reasonably good ROC curve, an AUC value of 0.825 and the highest hitrate: 84,3%. The variable's used in our final model after feature selection are reported in appendix A.

Comparison with the non-temporal model (option 0)
To compare the model with non-temporal variables to our model with temporal variables we look at the AUC values from table 8. We can conclude that the best result is achieved without the temporary variables, but with the C5.0 model with boosting applied. Hoogendoorn, M. et al achieved an AUC value of 0.834 using the CHAID decision tree, which is lower than the 0.898 that we achieved with C5.0 applied. Table 9 shows the confusion matrix of the model using the C5.0 algorithm and non-temporal variables.

Confusion Matrix		**Predicted**	
		No-CRC	CRC
Actual	No-CRC	41.694	3.538
	CRC	13	188

Table 9: The confusion matrix for the model without temporary variables

5. Discussion

The objective of the present study was to determine whether the number of consults per patient affected the chance to get colorectal cancer. In summary, it appears that the use of data mining allows to forecast colorectal cancer when taking the diagnoses, medication and referrals of every patient into account. This study makes several contributions to the literature. First, the model is applied to the dataset to describe the method and results. Second, a method to create a dataset that is viable for different models, risk models and applications.

The results showed the improvements in confusion matrix and hitrate, when the C5.0 algorithm is applied with the correct variables and options. The hitrate for CRC patients is significant higher in comparison with the standard options for the C5.0 node. Besides that, we achieve an improved results by using the machine learning techniques discussed in section 3.2. Table 4 and Figure 11 in the result section are showing the optimal model to predict CRC for patients in Utrecht. The C5.0 model is applied to specific data for CRC patients. However, this model appears to be very useful for large datasets and various analyses.

Closely related to colorectal cancer is gastric cancer. A study on diagnosis of gastric cancer shows that with the use of a decision tree an accuracy of 92,1% can be achieved (Su, et al., 2007). For these diagnoses, serum levels are measured. Their decision tree was generated using the Gini method with non-linear combinations. Furthermore, they identify nine serums of colorectal cancers as a control group and manage to predict seven as not being gastric cancer.

The limitations to the approach taken in this study largely relate to the amount of data. After preparing the data for modelling the number of classified patients was relatively low. Therefore, other industries may produce different results. Raghupathi W. and Raghupathi V. (Raghupathi & Raghupathi, 2014) summarize the different possibilities available in the area of Big Data and health care. They describe how techniques like Hadoop can be applied to enormous datasets like the U.S. healthcare system . However, they do not apply these techniques to different examples showing the behaviour of the data.

Besides that, the study has uncovered several important factors that are associated with the details of the healthcare sector. The ICPC diagnoses can be affected by the opinion of the GP. Moreover, patient feelings are not taken into account in this model. Results of this study may suggest a broader hypothesis for further research into these factors. This research can contain more data about patients, use more runtime and adjust the parameters and variables that have been used in this model.

6. References

Barzdins, G., Gosko, D., Rituma, L., & Paikens, P. (2014). Using C5.0 and Exhaustive Search for Boosting Frame-Semantic Parsing Accuracy. *In: Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pp. 4476-4482.

Bellaachia, A., & Guven, E. (2006, april 22). Predicting Breast Cancer Survivability Using Data Mining Techniques. *In: Proceedings of Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006)*.

Brissaud, J. (2005). The meanings of entropy. *Entropy 7*, 68-96.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780), 1612.

Hoogendoorn, M., Moons, L. M., Numans, M. E., & Sips, R.-J. (2014). Utilizing Data Mining for Predictive Modeling of Colorectal Cancer using Electronic Medical Records. *In: Proceedings of the 2014 Brain Informatics and Health Conference*.

Jemal, A., Siegel, R., Ma, J., & Zou, Z. (2014). Cancer statistics, 2014. *CA: a cancer journal for clinicians, 64*(1), 9-29.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In IJCAI (Vol. 14, No. 2, pp. 1137-1145)*.

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). *Supervised Machine Learning: A Review of Classification Techniques, Informatica.*

Kumar, H., Katiyar, P., Singh, A., & Rajkumar, B. (2014). Effect of different pruning severity on Growth and Yield of Ber (Zizyphus mauritiana Lamk). *International Journel of Current Microbiology and Applied Sciences, 3*(5), 935-940.

Liu, H., Li, J., & Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics Series*, 51-60.

Nowozin, S. (2012). Improved Information Gain Estimates for Decision Tree Induction.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning (Vol. 1).* San Mateo, CA: Morgan Kaufmann.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2*(1), 3.

Su, Y., Shen, J., Qian, H., Ma, H., Ji, J., Ma, H., . . . Shou, C. (2007). Diagnosis of gastric cancer using decision tree classification of mass spectral data. *Cancer science, 98*(1), 37-43.

Appendix A

This appendix is showing all fields used in the final model.

Rank	Field	Value	Rank	Field	Value
1	Age	1,0	36	ATC_Code_S_Q1	1,0
2	ATC_Code_B_Q4	1,0	37	Specialisme_interne_geneeskunde_Q4	0,999
3	ATC_Code_C_Q4	1,0	38	Specialisme_interne_geneeskunde_Q3	0,998
4	ICPC_B_Q4	1,0	39	ATC_Code_S_Q3	0,998
5	ATC_Q4	1,0	40	ATC_Code_L_Q2	0,997
6	ATC_Code_C_Q3	1,0	41	ATC_Code_S_Q4	0,996
7	ATC_Code_C_Q1	1,0	42	ICPC_S_Q1	0,996
8	ATC_Code_A_Q4	1,0	43	ATC_Code_M_Q3	0,994
9	ATC_Code_C_Q2	1,0	44	ICPC_D_Q3	0,99
10	Total_Spec_ICPC_ATC	1,0	45	ATC_Code_H_Q4	0,985
11	ATC_Code_B_Q3	1,0	46	ICPC_W_Q1	0,984
12	ICPC_K_Q4	1,0	47	ATC_Code_Y_Q4	0,984
13	ICPC_D_Q4	1,0	48	ATC_Code_H_Q2	0,981
14	ATC_Code_B_Q2	1,0	49	ATC_Code_M_Q1	0,981
15	ATC_Q3	1,0	50	ICPC_B_Q3	0,979
16	ICPC_Q4	1,0	51	ICPC_K_Q3	0,977
17	ATC_Code_B_Q1	1,0	52	ICPC_S_Q2	0,977
18	ATC_Q2	1,0	53	ICPC_H_Q2	0,973
19	Specialisme_longfunctie_lab_Q2	1,0	54	ATC_Code_L_Q1	0,972
20	Specialisme_overig_Q2	1,0	55	Specialisme_laboratorium_Q3	0,969
21	ICPC_T_Q4	1,0	56	ICPC_W_Q3	0,968
22	ATC_Q1	1,0	57	ICPC_A_Q1	0,966
23	ICPC_Slope	1,0	58	ICPC_D_Q2	0,962
24	ATC_Code_A_Q3	1,0	59	ICPC_U_Q2	0,958
25	ATC_Slope	1,0	60	ICPC_W_Q4	0,958
26	ICPC_A_Q4	1,0	61	ATC_Code_M_Q2	0,948
27	ATC_Code_A_Q1	1,0	62	ICPC_H_Q4	0,938
28	ICPC_K_Q1	1,0	63	ICPC_T_Q1	0,937
29	ICPC_U_Q4	1,0	64	ATC_Code_N_Q3	0,932
30	ATC_Code_A_Q2	1,0	65	ICPC_W_Q2	0,93
31	ATC_Code_S_Q2	1,0	66	ICPC_X_Q1	0,929
32	Specialisme_gastro_enterologie_Q4	1,0	67	ICPC_X_Q2	0,927
33	Specialisme_laboratorium_Q4	1,0	68	Specialisme_oogheelkunde_Q1	0,925
34	ATC_Code_M_Q4	1,0	69	ICPC_P_Q4	0,914
35	ATC_Code_N_Q4	1,0	70	Spec_Q4	0,903